Vector-Based Kernel Weighting: A Simple Estimator for Improving Precision and Bias of Average Treatment Effects in Multiple Treatment Settings

Melissa M. Garrido[1,2]
Jessica Lum[1]
Austin B. Frakt[1,2,3]
Steven D. Pizer[1,2]

1. Department of Veterans Affairs
2. Boston University School of Public Health, Boston, MA
3. Harvard University School of Public Health, Boston, MA

This version: July 31, 2018

This work was supported by VA HSR&D IIR 16-140 (PI Garrido). The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

Abstract

Treatment effect estimation must account for endogeneity, in which factors affect treatment assignment and outcomes simultaneously. By ignoring endogeneity, we risk concluding that a helpful treatment is not beneficial or that a treatment is safe when actually harmful. Propensity score matching or weighting adjusts for observed endogeneity, but matching becomes impracticable with multiple treatments, and weighting methods are sensitive to propensity score model misspecification in applied analyses. We used Monte Carlo simulations (1,000 replications) to examine sensitivity of multi-valued treatment inferences to propensity score weighting or matching strategies. We consider four variants of propensity score adjustment: inverse probability of treatment weights (IPTW), kernel weights, vector matching, and a new hybrid that is easily implemented – vector-based kernel weighting (VBKW). VBKW matches observations with similar propensity score vectors, assigning greater kernel weights to observations with similar probabilities within a given bandwidth. We varied degree of propensity score model misspecification, sample size, treatment effect heterogeneity, and sample distribution across treatment groups. Across simulations, VBKW performed equally or better than the other methods in terms of bias and efficiency. VBKW may be less sensitive to propensity score model misspecification than other methods used to account for endogeneity in multi-valued treatment analyses.

1. Introduction

Most propensity score guidance is restricted to methods for matching individuals with similar propensity scores across two groups (treatment, no treatment). Many treatments, however, have multiple levels. They may be continuous (drug dose) or categorical (several drug types). Consider, for example, a comparative effectiveness study that compares the ability of three interventions to prevent nursing home placement among individuals with functional and/or cognitive impairment: participant-directed home and community-based services, monthly home health aide services, and adult day care programs. If longitudinal data are not available, and if a valid instrumental variable is not available, a propensity score analysis that reduces bias due to observed confounding may be the best analytic strategy. In a traditional propensity score analysis, a series of dichotomous comparisons would be made: home and community-based services versus the other programs, home health aide services versus the other programs, home health aide services versus adult day care programs, and so on.

However, restricting treatments to binary indicators obscures between-group differences, including nonlinear relationships between treatment level and outcome (Cattaneo 2010). For instance, a binary comparison of adult day care versus the other two programs may not show a large difference in nursing home placement rates. However, if the three programs were to be compared simultaneously, one might find that adult day care is superior to home health aide services but not to participant-directed home and community-based services. Accounting for all values of a treatment variable in a single equation (rather than using several equations to make binary comparisons) helps ensure that a propensity score for a multi-valued treatment leads to treatment effect estimation among patients who have a non-zero chance of receiving any of the values of the treatment (i.e., that the assumption of common support is valid) (Rassen et al. 2013). For instance, in our example, a binary model of the probability of receiving participant-directed services will not distinguish between patients who have a chance of receiving adult day health care or home health aide services from patients who have a chance of receiving adult day health care but zero chance of receiving home health aide services. Inclusion of patients who have zero chance of receiving home health aide services in the analysis violates the assumption of common support necessary for propensity scores to reduce observed selection bias.

As the number of treatment groups increases, the option to estimate a single multinomial model, a generalized propensity score model, versus multiple binary models becomes more attractive. A generalized propensity score is the probability of receiving one treatment level, conditional on observed covariates. Each level is represented by a different propensity score (Imai & Van Dyk 2004; Imbens 2000).

The best way to use propensity scores for multinomial treatments is unknown; it is unclear when choice of a different weighting or matching strategy leads to divergent inferences. A popular method is inverse probability of treatment weights (IPTWs) based on the propensity score (e.g., McCaffrey, Griffin et al. 2013). However, IPTWs are sensitive to extreme values of the propensity score. If the extreme weights are caused by misspecification of the estimated propensity score model, use of IPTWs will lead to a biased estimate of the treatment effect. As the number of treatment groups increases, so does the likelihood of obtaining extreme values of the propensity score for at least one treatment group (Lopez & Gutman 2017).

Other options for incorporating generalized propensity scores into analyses include subclassification, matching, or kernel weights. Subclassification may not reduce selection bias as much as weighting when traditional numbers of strata (often 5) are used, and the optimal number of strata required to reduce selection bias may vary with sample size (Lunceford & Davidian 2004; Stuart 2010). Matching across groups becomes increasingly impractical, though not impossible (Rassen et al. 2013; Ratkovic 2012; Lopez and Gutman 2017), as the number of treatment groups increases; covariate distributions may not overlap well across multiple groups. Kernel weights, in which weights are assigned to comparison observations within a given bandwidth of the treated observation's propensity score, are less popular than IPTWs but minimize the influence of extreme weights (DiNardo & Tobias 2001; Garrido et al. 2014). However, there is little guidance to facilitate choice between IPTWs, matching, or kernel weighting.

In empirical analyses, investigators often encounter situations that may affect the ability of a propensity score to reduce selection bias. Nonlinearity in the data generating process (DGP) for the true propensity score, number of treatment groups, distribution of sample across treatment groups, and treatment effect heterogeneity will interact to influence bias and efficiency of treatment effect estimates. The degree to which different weighting or matching strategies lead to robust inferences in messy empirical analytic scenarios with multiple treatment groups is unknown.

Here, we examine the extent to which inferences are likely to diverge among four methods of matching or weighting on the propensity score. We examine IPTW and kernel weighting as well as one variant of matching (vector matching) that looks especially promising for reducing covariate imbalance in studies of treatments with multiple levels (Lopez and Gutman 2017). However, vector matching is relatively complex to implement. Finally, we introduce a hybrid of kernel weights and vector matching that is easier to implement, which we term vector-based kernel weighting (VBKW).

In the following sections, we describe treatment effects of interest, weighting and matching strategies in greater detail, reasons we might expect inferences to diverge based on choice of weighting or matching strategy, and our Monte Carlo simulation design. We show that VBKW is an easy-to-use weighting method that improves bias relative to existing methods across a wide range of true propensity scores and distributions of the sample across treatment groups.

2. Treatment effects of interest

We are interested in bias and efficiency of average treatment effects (ATEs) and average treatment effects on the treated (ATTs). Again, consider a treatment with three levels: A, B, and C, and let $E[Y_A]$ represent the estimated outcome when everyone in the sample receives level A. For this treatment, 3 ATEs ($E[Y_A] - E[Y_B]$; $E[Y_A] - E[Y_C]$; $E[Y_B]- E[Y_C]$) and 9 ATTs (each ATE evaluated among individuals who received a single treatment) can be estimated (McCaffrey, Griffin, et al. 2013). If the treatment effect is homogenous across the sample, the ATE and ATTs will be equal. In the more likely empirical case where treatment effectiveness varies by individual characteristics, the ATE and ATTs will not be equivalent. More formal definitions of our treatment effects of interest are presented in Appendix 1.

3. Weighting and Matching Methods

3.1 Inverse probability of treatment weights (IPTWs)

In IPTWs, observations receive weights equal to the inverse of their propensity scores (Hirano, Imbens, & Ridder 2003; Imbens 2004; see Appendix 1 for more details). Incorrectly estimated IPTWs may have extreme values, however, increasing treatment effect estimate variance (Stuart 2010). IPTWs permit average treatment effect (ATE) calculation but can be modified to calculate average treatment effects on the treated (ATTs) by assigning treated individuals a weight of one (called "weighting by the odds" or standardized mortality/morbidity ratio weighting) (Hirano et al. 2003; Ellis et al. 2013; Stuart, DuGoff, et al. 2013). We examined normalized IPTWs (Hirano et al. 2003). We do not consider IPTW adjustments such as trimming or truncating, as those methods employ arbitrary cut-points and may lead to estimates that are difficult to interpret (Harder, Stuart, & Anthony 2010; Stuart 2010; Lee, Lessler, & Stuart 2011).

As in all propensity score analyses, IPTW analyses are restricted to the range of common support. For IPTWs, this is often operationalized broadly as including observations between the maximum of the minima and the minimum of the maxima of each treatment group's propensity score (Caliendo & Kopeinig 2008). For instance, consider a treatment with three levels (A, B, C): 1) Find the maximum of the minima of the propensity score for treatment A (p(A)) across each treatment group, 2) Find the minimum of the maxima of p(A) across each treatment group, 3) Drop any observation with a value of p(A) outside of the region identified by steps 1 and 2, and 4) Repeat steps 1-3 for p(B) and p(C) (Lopez & Gutman 2017).

3.2 Kernel weights (KW)

KWs permit ATT calculation by assigning treated observations a weight of one and assigning weights to comparison observations within a given bandwidth of the treated observation's propensity score according to a kernel function (DiNardo & Tobias 2001). By doing this, KWs may have fewer extreme values than IPTWs. ATTs from different treatment groups can be combined to calculate ATEs. Treatment effect estimates from KWs are sensitive to the bandwidth, the range from a treated observation's propensity score, used to construct the weight (Caliendo & Kopeinig 2008). Smaller bandwidths lead to more exact matches, which reduces bias. However, if the bandwidth is too small, fewer observations will be included in the analytic sample and variance will increase. For that reason, we considered a bandwidth identified as optimal because it has been shown to minimize mean squared error (MSE) of the estimated propensity score model over the sample without sacrificing smoothness of the estimator: a constant of 0.06 (Heckman Ichimura & Todd 1997).[1] KWs are assigned according to kernel functions, where higher weights are given to comparison individuals with propensity scores most similar to treated individuals within the bandwidth. Weights are not as sensitive to choice of kernel function as they are to bandwidth (Caliendo & Kopeinig 2008), and we used the commonly used Epanechnikov kernel for all KW calculations (DiNardo & Tobias 2001; Busso, DiNardo, & McCrary 2009). The Epanechnikov kernel is the "optimal kernel" in that it minimizes MSE well at both interior and boundary points (Fan et al. 1997). Weights are normalized to sum to one in each treatment group (Imbens 2004; Busso et al. 2009).

[1] In future work, we will explore the use of a bandwidth that is dependent on the standard deviation of the logit of the propensity score in our KW.

KWs are composed from observations with similar probability of treatment and similar probability of non-treatment. This could be interpreted as including observations with similar probability of all treatment levels, or as including observations with similar probability of receiving each of the treatment levels being compared. For a binary treatment, these two interpretations are equivalent. However, consider a multiple treatment case where we are interested in the effect of treatment A vs treatment B (but where there are others who received treatment C). In this case, weights could be constructed among observations with similar probability of treatment A and similar probability of treatment B, within the range of common support (so all observations have a non-zero probability of treatment A, B, and C). Alternatively, weights could be constructed among observations with similar probability of treatment A, similar probability of treatment B, **and** similar probability of treatment C (Lopez and Gutman 2017). We implement the first, broader, option in our KW construction. The second option relies on identification of similar vectors of propensity scores, which is part of vector matching (Lopez and Gutman 2017) and which we add to traditional kernel weight construction to create a new hybrid strategy, vector-based kernel weighting (VBKW).

3.3 Vector matching (VM)

As its name suggests, VM creates matches after identifying vectors of similar propensity scores across groups (Lopez and Gutman 2017). VM lends itself to the calculation of an ATT, but ATT estimates can be combined to create ATEs. Lopez and Gutman recently developed this procedure and found that it leads to better matches (lowest bias in covariate distribution among treatment groups) than common referent matching or IPTWs (2017). Matching within vectors ensures treatment effect estimates are applicable to observations with similar probability of receiving any of the treatments. However, this process becomes more difficult to implement as number of treatment groups increases, which makes it less likely that there will be available matches.

VM identifies similar vectors of propensity scores two ways when creating a matched set: first, by clustering, and then, by 1:1 greedy matching with replacement (Lopez and Gutman 2017). The ability to create good matches relies on several steps. After dropping observations outside of the range of common support, Lopez and Gutman recommend refitting the propensity score model and using k-means clustering on the logit of the propensity score for each treatment group to create strata with similar vectors of propensity scores. Clustering and matching is repeated as many times as there are treatment groups (3 rounds of clustering and matching for a treatment with 3 groups). For a treatment with values A, B, and C, treatment group A serves as the first reference group. Within strata formed from clusters of observations with similar values of the logit of p(C), observations from treatment groups A and B are matched based on values of the logit of p(A). Matches occur within a caliper of 0.25*SD(logit(p(A))). Then, within strata formed from clusters of observations with similar values of the logit of p(B), observations from treatment groups A and C are matched based on values of the logit of p(A). The observations from treatment group A that matched to observations in both other treatment groups, as well as the matches from groups B and C are retained. Similar steps are repeated with treatment group B and treatment group C as the reference groups. VM creates different matched samples, depending on the treatment effects of interest. For instance, when the reference group is A, we can calculate the ATT of A vs B and the ATT of A vs C, among observations with treatment A.

3.4 Vector based kernel weighting (VBKW)

In VBKW, weights are assigned based on a kernel function (as described above), but the weights are assigned to observations that have similar vectors of propensity scores for each treatment. VBKW includes elements of VM and kernel weighting, but it is simpler to implement than VM. No clustering or matching steps are required. Recall that in traditional kernel weighting, weights for an observation in treatment group A equal one, and nonzero weights for observations in treatment group B are assigned if their value of p(A) is within a bandwidth of .06 of the treated observation's value of p(A). Values of p(B) and p(C) are not considered. In contrast, in VBKW, to ensure that weights are created for observations with similar vectors of propensity scores, we assign nonzero weights to observations in treatment group B if their value of p(A) is within a bandwidth of .06 of the treated observation's value of p(A), **and** if their value of p(B) is within a bandwidth of .06 of the treated observation's value of p(B), **and** if their value of p(C) is within a bandwidth of .06 of the treated observation's value of p(C). This means that nonzero weights are assigned to controls with a similar propensity score **vector** instead of just being similar on p(A). Rather than creating several subsets of matches, as in VM, VBKW creates one single subpopulation, allowing for easier comparison of estimated treatment effects. As is the case in kernel weighting and VM, ATT estimates from VBKW can be combined to form ATEs.

4. Reasons we might expect inferences to diverge based on choice of weighting or matching strategy

As sample sizes approach infinity, results from any propensity score method should converge, but researchers need guidance for their use in finite empirical samples. We wish to identify scenarios in which inferences are most likely to diverge. For instance, we expect estimates from kernel weights (low emphasis on extreme weights) to be less biased than estimates from IPTWs when the true data generating process for the propensity score is nonlinear. We expect differences in inferences to be more likely when the presence of extreme weights is more likely or when identification of matches may be more difficult. We expect this to occur when: 1) the true propensity score includes more nonlinearity and nonadditivity, 2) the number of treatment groups increases, 3) the sample size decreases, 4) the sample is distributed more unevenly across treatment groups, 5) when there are heterogeneous treatment effects, and 6) when there is greater pre-weighting imbalance in observed covariates across groups[2] (Lee, Lessler, & Stuart 2010; Lee et al. 2011; Rassen et al. 2013; Setoguchi et al. 2008).

5. Methods

We used simulated datasets to understand the relative ability of different propensity score matching and weighting strategies to reduce selection bias in treatment effect estimation. We ranked weighting and matching strategies' ability to produce unbiased treatment effects.

We began with a simulation design with a known data generating process so that we can isolate and identify the influence of weighting/estimation strategies and analytic scenarios on treatment effect estimates. We based our definitions of true propensity scores on an established simulation protocol where covariates are correlated and the true propensity score exhibits varying degrees of nonlinearity and nonadditivity (Table 1) (Lee et al. 2010; Wyss et al. 2014; Lee et al. 2011;

[2] Simulations in which pre-weighting imbalance across groups is varied are planned but not yet completed.

Setoguchi et al. 2008; Austin 2012). In future studies, our simulation design will be extended to include plasmode simulations of empirical data, which will enable us to verify that the simulation results are not due to our choice of data generating processes.

The initial true propensity score model was a multinomial logistic model of covariates (X_1-X_7), where X_1-X_4 were confounders and X_5-X_7 were associated with treatment only (Setoguchi 2008). All covariates except for X_2, X_7, and X_{10} were drawn from the standard normal distribution. Covariates X_2 and X_{10} were drawn from a normal distribution with mean 1 and standard deviation 1. Covariate X_7 was drawn from a normal distribution with a mean of -1 and standard deviation of 1. Our outcome was a linear function of X_1-X_4, X_8-X_{10}, and treatment assignment (Lee et al. 2010). The true ATEs, $E[Y_A] - E[Y_B]$, $E[Y_A] - E[Y_c]$, and $E[Y_B] - E[Y_C]$, were set to have values of: -0.1, -0.2, and -0.1, respectively. The true ATTs were equal to the true ATEs when treatment effects were homogenous. Following Setoguchi et al. 2008's protocol, we set X_1 and X_5 to have a correlation coefficient of 0.2, and X_3 and X_8 to have a correlation coefficient of 0.2. In addition, both X_2 and X_6 were set to have a correlation coefficient of 0.9, as were X_4 and X_9. After setting the correlation coefficients, X_1, X_3, X_5, X_6, X_8, X_9 were dichotomized ($X_{1new} = 0$ if $X_1 \leq \overline{X}_1$, $X_{1new} = 1$ if $X_1 > \overline{X}_1$).

To generate the true propensity score and create treatment groups, we calculated a multinomial logit model:

$$p(A) = \frac{1}{1 + e^{X\beta_B} + e^{X\beta_C}}$$

$$p(B) = \frac{e^{X\beta_B}}{1 + e^{X\beta_B} + e^{X\beta_C}}$$

$$p(C) = \frac{e^{X\beta_C}}{1 + e^{x\beta_B} + e^{x\beta_C}}$$

where $X\beta_B = .2*(-.2 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7)$

and $X\beta_C = -.9*(-.1 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7)$

Coefficients in this model were selected to reflect magnitudes often present in empirical analyses, and that were similar to those used in other simulation studies involving treatments with multiple levels. We tested sensitivity of results when the coefficient magnitudes were increased from .2 to 1.2 (relative risk ratio (RRR) from 1.22 to 3.32) and from -.9 to -1.9 (RRR of .4 to .15) in the linear predictor functions of treatments B and C, respectively. Results were qualitatively unchanged.

To assign observations to treatment groups, we generated a random number from the uniform distribution that represents probability of treatment (denoted by j). Observations were assigned to

treatment group A if $j < p(A)$, to treatment group B if $p(A) \leq j < p(A) + p(B)$, and to treatment group C if $j \geq p(A) + p(B)$. In order for us to vary distribution of the sample across treatment groups (Table 1, line 3), we began with a sample size of 100,000. After generating treatment group assignment for the entire simulated dataset, we randomly drew observations from each treatment group. For instance, in our simulations of n=999 with equal treatment distribution across three treatment groups, we randomly drew 333 observations from treatment groups A, B, and C.

For each true propensity score (all possible combinations of Table 1 characteristics), we used each weighting and matching strategy to estimate all possible ATEs and ATTs in sample sizes expected to occur in prospective observational cohort studies (n=999), and in administrative data analyses (n=9,999) of empirical health services research questions. Each simulation consisted of 1,000 replications. Analyses were conducted in Stata version 14 (StataCorp 2015).

Here, estimated propensity scores were calculated using maximum likelihood estimation (multinomial logit regression) on the main effects of X_1-X_4 and X_8-X_{10}. Differences in inferences from weighting strategies when other estimation strategies are employed are left for future work (covariate balancing propensity scores [Imai & Ratkovic 2014], generalized boosting methods [McCaffrey, Griffin, et al 2013]).

5.1 Outcomes

When varying simulations by 7 propensity score model misspecifications, 12 estimands, 3 types of sample distributions across treatment groups, and 3 types of treatment effect distribution, (7 x 12 x 3 x 3), we obtain 756 unique analytic scenarios from which to compare performance of our estimators. Across each scenario, we evaluated bias and efficiency. For each ATE and ATT, we measured bias (distance between the true treatment effect and mean estimated treatment effect, where smaller distances represent less bias). We also calculated percent bias (bias as percent of standard deviation) (Kang & Schafer 2007). Efficiency of each treatment effect estimate was evaluated by interquartile range (IQR) magnitude, root-mean-squared error (RMSE), and mean absolute error (MAE). We counted the number of analytic scenarios in which each matching and weighting strategy produced estimates where bias was < 40% of the standard deviation, an indication of situations in which test statistics are still likely to perform well (Kang & Schafer 2007). Among those scenarios, we ranked matching and weighting strategies by efficiency.

6. Results

We focus here on results from simulations where n=999 and the number of treatments (denoted by k) =3. Preliminary results from simulations where n=9,999 and k=3, and from simulations where n=999 and k=4, show similar patterns of bias and efficiency as those we report here. The magnitude of differences in bias and efficiency as sample size or number of treatment groups increases will be described as the study progresses.

6.1 Overall performance of IPTW, KW, VM, and VBKW

When n=999 and k=3, estimates based on IPTW were more likely to be biased and inefficient than estimates based on KW, VM, or VBKW (Table 2). Of the 756 analytic scenarios we ran for IPTW, only 221 (29.2%) produced estimates where bias was less than 40% of the estimate's standard deviation. VBKW-based estimates, in contrast, are the most likely to be unbiased and

efficient. VBKW produced estimates where bias was less than 40% of the estimate's standard deviation in 73.3% of the analytic scenarios.

Across all estimates (Table 2) and across estimates with <40% bias (Table 3), VBKW produced the most efficient estimates (median RMSE among estimates with <40% bias = .058, median RMSE from other strategies ranged from .066 - .080).

6.2 Sensitivity to misspecification of the propensity score model

Regardless of weight type, the least biased and most efficient estimates were observed when the true propensity score included only additive main effects. As expected, bias was greatest when the true propensity score included moderate nonlinearity and nonadditivity (scenario G) that were not captured in the estimated propensity score model.

IPTW produces similar results to the other techniques in ideal settings (true propensity score only includes additive main effects, homogenous treatment effect, equal distribution of sample across treatment groups), but it is more sensitive to propensity score model misspecification than any of the other methods we examined (Figure 1; Appendix 2). The median absolute bias of estimates produced by VBKW is less sensitive to propensity score misspecification than in IPTW, KW, or VM (Figure 2).

6.3 Sensitivity to distribution of sample across treatment groups

Again, VBKW estimates were the least sensitive (smallest changes in percent bias, as well as the lowest overall percent bias) to variations in the distribution of the sample across treatment groups (Figure 3). The magnitude of IQRs changed similarly across all weighting and matching strategies as treatment distribution became more skewed.

6.4 Sensitivity to treatment effect heterogeneity

In the presence of homogenous treatment effects and heterogeneity due to a non-confounding variable, both VBKW and VM were more likely to produce estimates with bias < 40% of standard deviation than IPTW or KW (details for heterogeneous treatment effects available from authors). As expected, in the presence of heterogeneity due to a confounder, all strategies were likely to produce biased estimates of the ATEs.

6.5 Performance across different estimands

When treatment effects were homogeneous or heterogeneous with respect to a nonconfounding variable, VBKW and VM were more likely to produce estimates with bias < 40% than IPTW or kernel weights, regardless of the estimand of interest (see Figure 3 for homogeneous effects). In turn, this allows VBKW and VM to produce less biased estimates of transitive treatment effects (i.e., calculating the ATT of B vs C among observations receiving treatment A from the ATTs of A vs B and A vs C among observations receiving treatment A).

7. Discussion

We investigated bias and efficiency of IPTW, KW, VM, and VBKW in analytic scenarios likely to be encountered in empirical analyses: misspecified estimated propensity score models, treatment effect heterogeneity, and sample distribution across treatment groups. The commonly

used IPTW strategy led to biased estimates more often than any other strategy we investigated. In nearly all scenarios, VBKW led to the least biased and most efficient estimates of the true treatment effect. These results suggest that VBKW may be less sensitive to propensity score model misspecification and sample distribution across treatment groups than other methods used to account for endogeneity in multi-valued treatment analyses. VBKW's performance was only slightly better than that of VM, but it is simpler to implement.

Estimates based on IPTW were especially sensitive to degree of propensity score model misspecification and skewed distribution of the sample across treatment groups. In addition, neither KW nor IPTW estimates appear well-suited to produce unbiased estimates of transitive ATTs. Transitive ATT estimates are less likely to be biased when weights are constructed among observations with similar vectors of propensity scores (VBKW, VM) than when they are constructed among observations within a range of common support defined by the maximum of minima and minimum of maxima of propensity scores (IPTW, KW) (Lopez and Gutman 2017).

7.1 Limitations

These results, while promising, need to be evaluated in light of several limitations. First, our simulations were based on an imposed DGP rather than an empirical one, potentially inaccurately reflecting scenarios likely to be encountered in applied analyses. In addition, we will verify our results with plasmode simulations based on DGPs present in empirical data (Franklin et al. 2014; see more details below). However, we obtained similar results when we used alternate coefficients in treatment models and alternate ways of generating treatment groups.

In addition, we deliberately estimated a misspecified propensity score model with only main effects. A well-done propensity score analysis should ensure that the propensity score is leading to adequate covariate balance (Garrido et al. 2014), but we wanted to understand the degree to which results are robust to misspecification (and to potential lapses in analytic quality). This follows the pattern of previous propensity score simulation studies (e.g., Setoguchi et al. 2008).

Relatedly, we do not test sensitivity of results to observed covariate choice or covariate measurement errors, nor do we test performance when propensity scores are combined with covariates in doubly-robust estimates. These important factors may affect estimates' bias and efficiency in finite samples (Stuart et al. 2010; Kang and Schafer 2007; McCaffrey, Lockwood, & Setodji 2013; Pearl 2009; Shadish 2013; VanderWeele & Arah 2011) and should be addressed after we have a better understanding of the relative performance of weighting and matching strategies for a given set of confounders.

7.2 Future directions

Future work will allow us to verify our results in simulations based on DGPs present in empirical data (plasmode simulations). A plasmode is a dataset based on empirical data generating processes that "has been constructed so that at least some aspect of the 'truth' of the data generating process is known" (Vaughan et al. 2009). Plasmode simulations were developed for genome and microarray research and are now being applied to electronic health data (Franklin et al. 2014; Franklin et al. 2017). Traditional simulations are often criticized for their artificiality, and empirical data analyses are limited by analysts' inability to observe the true treatment effect.

Plasmode simulations overcome these limitations by combining the benefits of a traditional simulation (known treatment effect that enables calculation of bias in treatment effect estimates) with the benefits of empirical data (empirical values of covariates and relationships among covariates are preserved). Plasmode simulations have the benefit of being derived from an empirical DGP while allowing us to observe a true treatment effect and thus the degree to which each weighting or matching strategy leads to biased treatment effect estimates.

In future work, we will also determine the degree to which inferences diverge when we use nonparametric and semiparametric methods of estimating propensity scores. Values of propensity score weights vary with propensity score estimation method. As a result, treatment effect estimates obtained after propensity score weighting are sensitive to propensity score estimation method; this is well-documented in studies of binary treatments (Harder et al. 2010; Stuart 2010; Imai & Ratkovic 2014; Kang & Schafer 2007). Covariate balancing propensity scores (estimated with generalized method of moments) and propensity scores created through generalized boosting methods rely less on investigator trial and error than maximum likelihood estimation methods to create a propensity score that achieves covariate balance across treatment groups (Harder et al. 2010; Garrido et al. 2014; Dehejia & Wahba 1999).

We may also be able to improve the performance of VBKW through adjustments to the bandwidth and by basing weights on the logit of the propensity score rather than on the propensity score itself (Stuart 2010). In addition, we will investigate the degree to which our patterns of results are robust to degree of pre-weighting imbalance across groups.

In order to develop useful guidance for empirical analyses, where bias cannot be ascertained, future work will compute a measure of covariate balance across treatment groups in the simulated data (Harder et al. 2010). We will verify whether the patterns of relative performance across weighting or matching strategy we observed in these analyses are similar for covariate balance.

Future work will also consider a measure of robustness to residual confounding. A limitation of propensity scores is that they only adjust for observed, not unobserved, confounding. To that end, we will identify how much unobserved confounding would need to be present for each strategy in each simulation scenario in order for the inference from the analysis to change. For each simulation scenario, we will rank the ATTs and ATEs produced by each strategy by degree of robustness to unobserved confounding. We will do this by following a weighted adaption of Rosenbaum's sensitivity analysis methods. (Liu, Kuramoto, & Stuart 2013; Rosenbaum 2002). For each treatment effect estimate, we will identify the smallest amount of unobserved selection bias that would need to be present to change the inference from rejection to acceptance of the null hypothesis of no treatment effect. Strategies that require relationships between an unobserved confounder and the treatment and between an unobserved confounder and the outcome to be stronger before inferences change are considered more robust.

8. Conclusion

When propensity scores are used in analyses of binary treatments, vector matching and weighting are implicitly conducted. Matching on the probability of being treated leads to matching on the probability of not being treated. If a treatment has more than two values, vectors

need to be explicitly included in the creation of propensity score matches or weights. If they are not included, the propensity score for only one treatment group will be balanced, and estimates are likely to be biased and inefficient. VM and VBKW both lead to less biased and more efficient estimates than IPTW or KW that do not include vectors when there are more than two treatment groups. VBKW is relatively simple to implement and creates a single weighted subpopulation, facilitating comparisons of ATTs and ATEs among observations eligible to receive any of the treatments under consideration.

9. References

Austin PC. Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based G-computation. Multivariate Behavioral Research 2012; 47(1): 115-135.

Busso M, DiNardo J, McCrary J. 2009. New evidence on the finite sample properties of propensity score reweighting and matching estimators [Discussion Paper]. Institute for the Study of Labor Discussion Papers No. 3998. 2009.

Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys 2008; 22(1): 31–72.

Cattaneo M. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. J Econometrics 2010; 155: 138-154.

Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association 1999; 94(448): 1053-1062.

DiNardo J, Tobias JL. Nonparametric density and regression estimation. Journal of Economic Perspectives 2001; 15(4):11–28.

Ellis AR, Dusetzina SB, Hansen RA, Gaynes BN, Farley JF, Stürmer T. Investigating differences in treatment effect estimates between propensity score matching and weighting: A demonstration using STAR*D trial data. Pharmacoepidemiology and Drug Safety 2013; 22: 138-144.

Fan J, Gasser T, Gubels I, Brockmann M, Engel J. Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. Ann Inst Statist Math 1997; 49(1): 79-99.

Franklin JM, Schneeweiss S, Polinksi JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Computational Statistics and Data Analysis 2014; 72: 219-226.

Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. Stat Med 2017; 36(12): 1946-1963.

Garrido MM, Kelley AS, Paris J, Roza K, Meier DE, Morrison RS, Aldridge MD. Methods for constructing and assessing propensity scores. Health Serv Res 2014; 49(5): 1701-1720.

Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological Methods 2010; 15(3): 234-249.

Heckman JJ, Ichimura H, Todd PE. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 1997; 64: 605–54.

Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 2003; 71(4): 1161-1189.

Imai K, Ratkovic M. Covariate balancing propensity score. J R Statist. Soc. B 2014; 76(1): 243-246.

Imai K, van Dyk DA. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association 2004; 99(467): 854-866.

Imbens GW. The role of the propensity score in estimating dose-response functions. Biometrika 2000; 87(3): 706-710.

Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. The Review of Economics and Statistics 2004; 86(1): 4-29.

Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 2007; 22(4): 523-539.

Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in Medicine 2010; 29: 337-346.

Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLOS One 2011; 6(3): e18174.

Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. Prev Sci 2013; 14: 570-580.

Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: A review and new ideas. Statistical Science 2017; 32(3): 432-454.

Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine 2004; 23: 2937-2960.

McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med 2013; 32: 3388-3414.

McCaffrey DF, Lockwood JR, Setodji CM. Inverse probability weighting with error-prone covariates. Biometrika 2013; 100(3): 671-680.

Pearl J. Remarks on the method of propensity score [Letter to Editor]. Statistics in Medicine 2009; 28: 1415-1424.

Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. Epidemiology 2013; 24: 401-409.

Ratkovic M. Identifying the largest balanced subset of data under general treatment regimes. Working paper [online]. Princeton University 2012.

Rosenbaum PR. Observational studies (2nd ed.). New York: Springer; 2002

Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. Pharmacoepidemiology and drug safety 2008; 17: 546-555.

Shadish WR. Propensity score analysis: Promise, reality, and irrational exuberance. J Exp Criminol 2013; 9: 129-144.

StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical Science 2010; 25(1): 1-21.

Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using electronic health data: Challenges and (some) solutions. eGEMS (Generating Evidence & Methods to Improve Patient Outcomes) 2013; 1(3): Article 4.

VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology 2011; 22: 42-52.

Vaughan LK, Divers J, Padilla MA, Redden DT, Tiwari HK, Pomp D, Allison DA. The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. Computational Statistics & Data Analysis 2009; 53: 1755-1766

Wyss R, Ellis AR, Brookhart MA, Girman CJ, Funk MJ, LoCasale R, Stürmer T. The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bCART, and the covariate-balancing propensity score. Am J Epidemiology 2014; 180(6): 645-655.

Table 1. Characteristics varied in Monte Carlo simulation and rationale for inclusion

Characteristic: Rationale	**Possible Levels**
1) Nonlinearity/nonadditivity in covariates included in propensity score (Lee et al. 2010; Lee et al. 2011; Setoguchi et al. 2008): To enable comparisons with existing research, we use Setoguchi et al.'s seven scenarios for true propensity scores that include various levels of nonlinearity (polynomials of covariates) and nonadditivity (interaction terms between covariates) (Setoguchi et al. 2008). As the model increases in nonlinearity and nonadditivity, we expect more disparate inferences (Setoguchi et al. 2008).	True propensity score is function of the following terms: a) $X_1,\ldots X_{10}$ b) $X_1,\ldots X_{10,}\ X_2^2$ c) $X_1,\ldots X_{10,}\ X_2^{2,}\ X_4^{2,}\ X_7^2$ d) $X_1,\ldots X_{10}$, X_1X_3, X_2X_4, X_4X_5, X_5X_6 e) $X_1,\ldots X_{10}$, X_1X_3, X_2X_4, X_4X_5, X_5X_6, X_2^2 f) $X_1,\ldots X_{10}$, X_1X_3, X_2X_4, X_4X_5, X_5X_6, X_5X_7, X_1X_6, X_2X_3, X_3X_4 g) $X_1,\ldots X_{10}$, X_1X_3, X_2X_4, X_4X_5, X_5X_6, X_5X_7, X_1X_6, X_2X_3, $X_3X_{4,}\ X_2^{2,}\ X_4^{2,}\ X_7^2$
2) Number of treatment groups: Evaluate performance in common numbers of treatment groups expected in empirical research (e.g., k=4: drug A + placebo, drug B + placebo, drugs A+B, placebo). We expect CBPS and kernel weights (less influence of extreme weights) to lead to the least biased estimates. We expect differences in inferences across strategies to be more likely as *k* increases (making covariate balance more difficult).	k = 3, 4
3) Distribution of sample across treatment groups (Rassen et al. 2013): Understand how well strategies approximate counterfactuals for treated individuals when there is relatively little information from comparison individuals. Unequal treatment group sizes often occur empirically. As group size decreases, weights will be constructed from fewer observations and will have greater variance. As the distribution becomes more skewed, we expect inferences from estimates based on non-stabilized IPTW (greater sensitivity to variance in weights) to diverge more than inferences from estimates based on other strategies.	a) Equal split across groups b) One treated group = 50% of observations, other groups split the remaining 50% equally c) One treated group = 10% of observations, other groups split the remaining 90% equally
4) Treatment effect heterogeneity: Understand how well strategies reduce bias in ATTs when ATTs are not expected to equal the ATE. We expect that differences in inferences that will arise with characteristics in rows 1-3 will be exacerbated in the presence of treatment effect heterogeneity (Rassen et al. 2013).[3]	Coefficient on treatment variable in outcome equation is: a) Constant (c) b) $c*X_{10}$ (associated with outcome) c) $c*X_2$ (confounder)

[3] Because one of the goals of this work is to develop practical guidance for applied investigators, we include a scenario that is likely to occur empirically, where heterogeneity is due to values of a confounder (4c in Table 1). There is no reason to expect any propensity score method would produce unbiased estimates of the ATE in this case, but we list it here for the sake of completeness.

Table 2. Summary of bias and efficiency of estimates across all analytic scenarios[a]

Weighting or matching strategy	Total number of analytic scenarios	Number (%) of analytic scenarios with <40% bias	Median bias as % of SD	Median absolute bias	Median IQR	Median RMSE	Median MAE
IPTW	756	221 (29%)	69.626	0.051	0.095	0.095	0.062
KW	756	356 (47%)	45.102	0.030	0.085	0.086	0.060
VM	756	542 (72%)	26.362	0.018	0.103	0.085	0.056
VBKW	756	554 (73%)	17.509	0.010	0.075	0.062	0.042

IQR = Interquartile range, RMSE = root-mean-squared error, MAE = mean absolute error, IPTW = inverse probability of treatment weights, KW = kernel weights, SD = standard deviation, VM = vector matching, VBKW = vector-based kernel weights

a) One analytic scenario is one combination of the elements from Table 1 and one treatment effect estimate (e.g., one analytic scenario includes n=999, true propensity score includes mild nonlinearity, k = 3, evenly split sample distribution across treatment groups, a homogeneous treatment effect, and an estimate of the ATT of A vs B among observations receiving A)

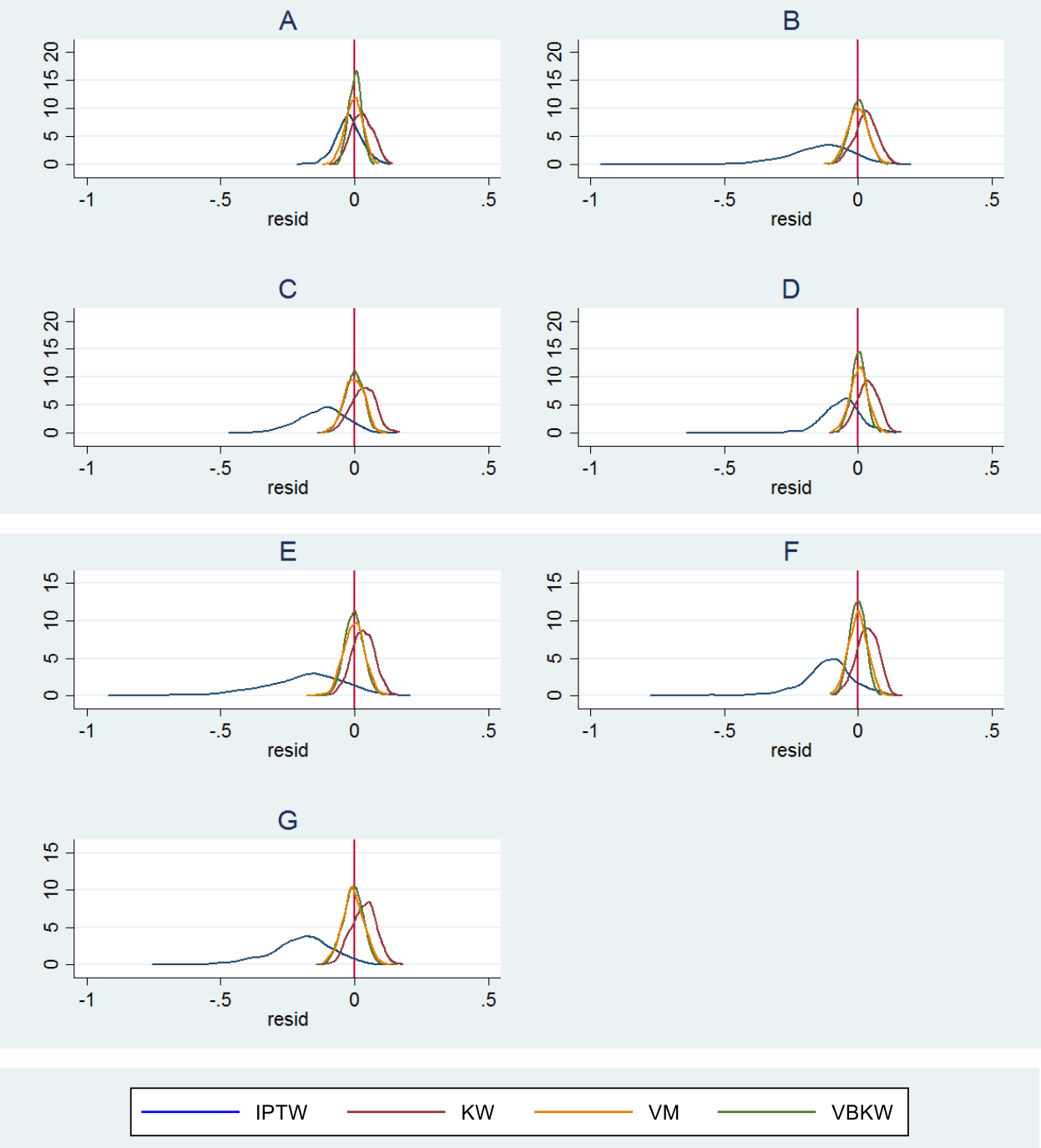
Table 3. Summary of bias and efficiency of estimates across analytic scenarios[a] that lead to <40% bias

Weighting or matching strategy	Number of analytic scenarios with <40% bias	Median bias as % of SD	Median absolute bias	Median IQR	Median RMSE	Median MAE
IPTW	221	19.935	0.014	0.083	0.066	0.042
KW	356	11.934	0.008	0.104	0.080	0.053
VM	542	17.186	0.012	0.103	0.080	0.053
VBKW	554	10.702	0.006	0.075	0.058	0.038

IQR = Interquartile range, RMSE = root-mean-squared error, MAE = mean absolute error, IPTW = inverse probability of treatment weights, KW = kernel weights, SD = standard deviation, VM = vector matching, VBKW = vector-based kernel weights

a) One analytic scenario is one combination of the elements from Table 1 and one treatment effect estimate (e.g., one analytic scenario includes n=999, true propensity score includes mild nonlinearity, k = 3, evenly split sample distribution across treatment groups, a homogeneous treatment effect, and an estimate of the ATT of A vs B among observations receiving A)

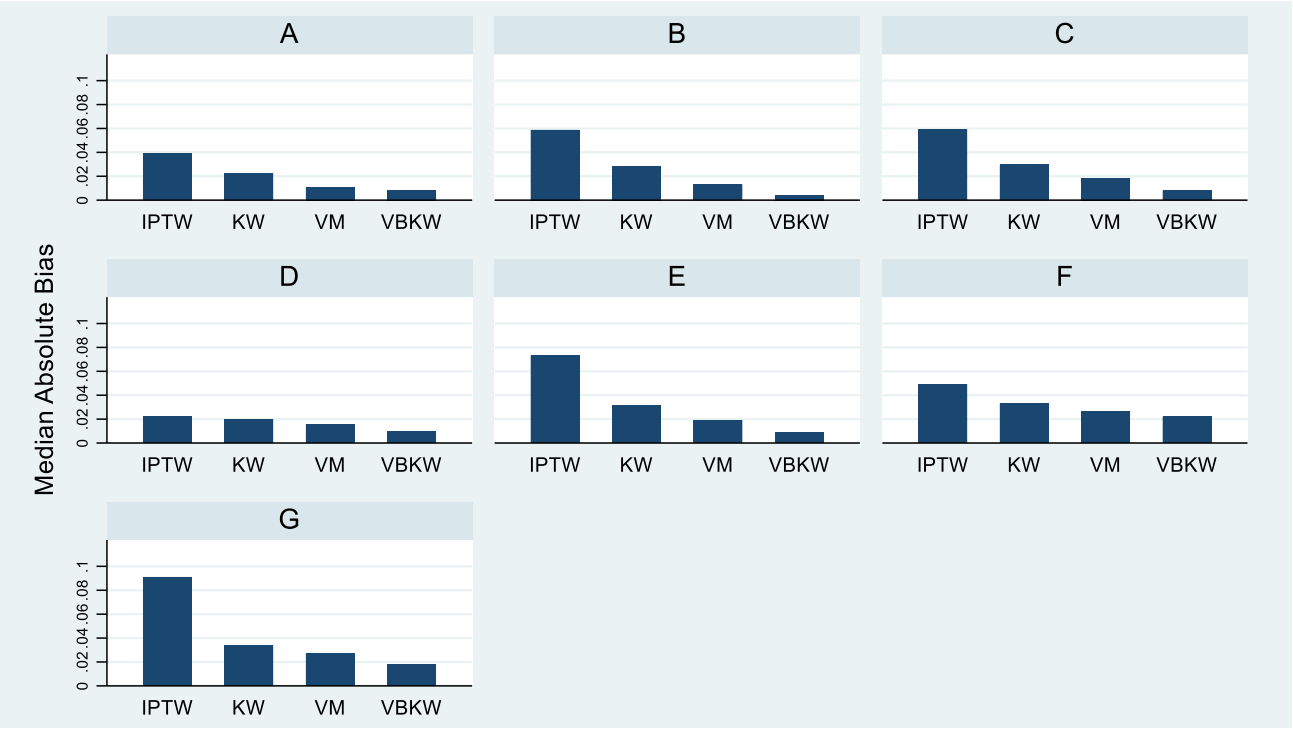
Figure 1. IPTW estimates are more likely to be biased and inefficient in the presence of propensity score model misspecification (panels B-G) than estimates from KW, VM, or VBKW.



IPTW = inverse probability of treatment weights, KW = kernel weights, VBKW = vector-based kernel weights, VM = vector matching.

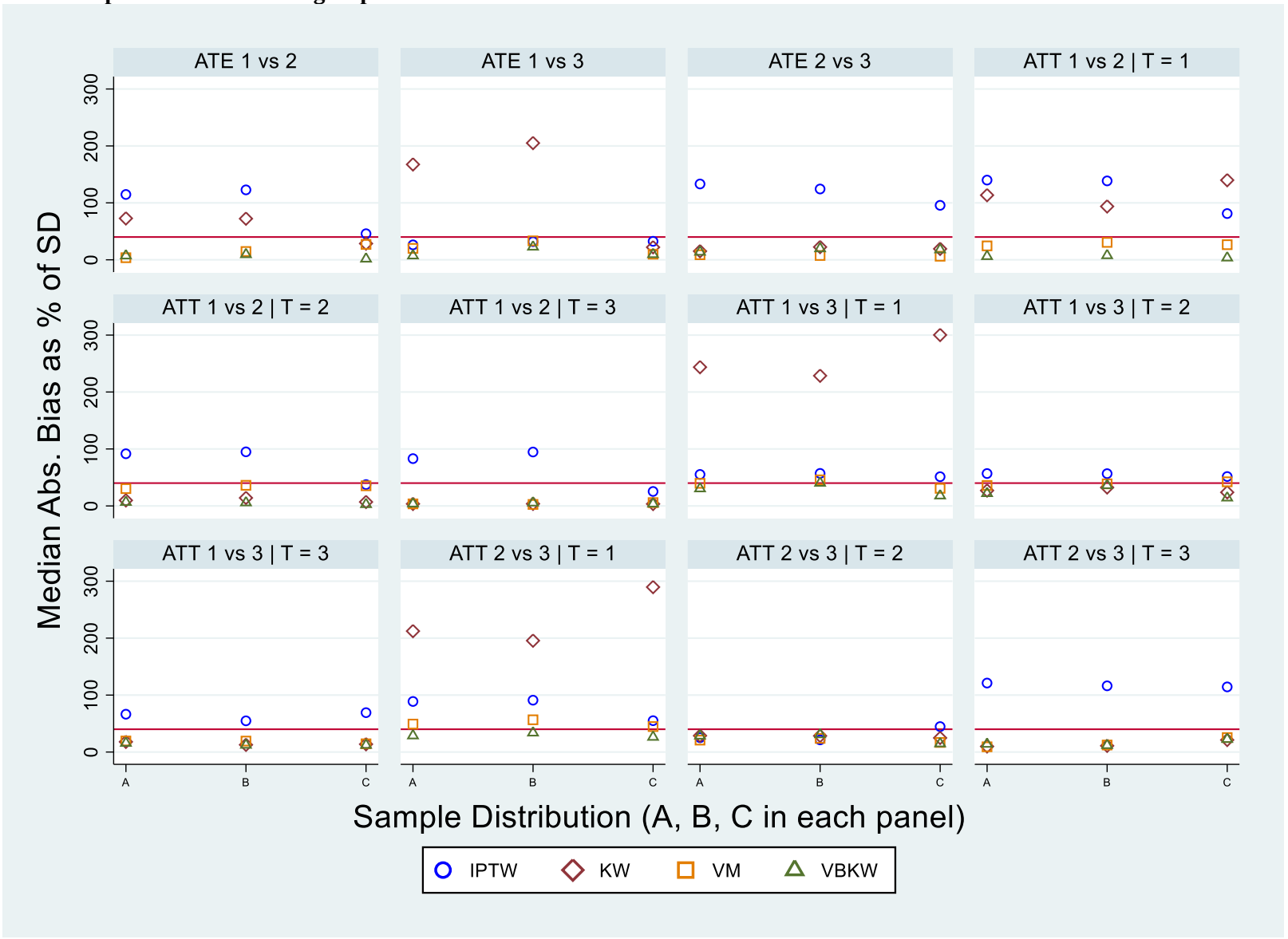
Density plot of residuals from ATE of treatment 1 vs 2 using IPTW, KW, VM, and VBKW with homogenous treatment effects, even distribution of sample across three treatment groups, and n=999. Estimated propensity score includes main effects only. True propensity score includes A) main effects only, B) mild nonlinearity, C) moderate nonlinearity, D) mild non-additivity, E) mild nonlinearity and mild non-additivity, F) moderate non-additivity, G) moderate nonlinearity and moderate non-additivity (Setoguchi et al. 2008) (see Table 1 for more details).

Figure 2. The median absolute bias of estimates produced by VBKW is less sensitive to propensity score misspecification (panels B-G) than in IPTW, KW, or VM.



IPTW = inverse probability of treatment weights, KW = kernel weights, VBKW = vector-based kernel weights, VM = vector matching.
Median absolute bias across all analytic scenarios, stratified by true propensity score (A-G). Estimated propensity score includes main effects only. True propensity score includes A) main effects only, B) mild nonlinearity, C) moderate nonlinearity, D) mild non-additivity, E) mild nonlinearity and mild non-additivity, F) moderate non-additivity, G) moderate nonlinearity and moderate non-additivity (Setoguchi et al. 2008) (see Table 1 for more details).

Figure 3. VBKW estimates were the least sensitive (smallest changes in percent bias, as well as the lowest overall percent bias) to variations in the distribution of the sample across treatment groups.



IPTW = inverse probability of treatment weights, KW = kernel weights, VBKW = vector-based kernel weights, VM = vector matching.
Median bias as a percent of standard deviation across scenarios with a homogenous treatment effect, stratified by estimand and sample distribution (A = equally distributed; B = one treated group has 50% of observations, other groups split the remaining 50% equally; C = one treated group has 10% of observations, other groups split the remaining 90% equally). Red lines indicate bias of 40% of the estimate's standard deviation.

Appendix 1

Setup and notation:

In a standard cross-sectional setting, we observe a sample of individuals i = 1, 2, . . ., N from a population.

Our sample size N is the sum of each the treatment group sizes: $N = N_1 + N_2 + \ldots + N_Z$.

Each individual has been assigned one of z possible treatment levels, where $z = 1, 2, \ldots, Z$. We observe the outcome variable y_i, the observed treatment level t_i, and a k_x x 1 vector covariates, $\mathbf{x}_i$. We also define an indicator variable $d_i(z) = 1(t_i=z)$ which is equal to 1 if unit i received treatment z and equal to 0 otherwise. We distinguish between the observed outcome y_i and the Z potential outcomes, $y_i(z)$. The observed outcome is given by

$y_i = d_i(1)y_i(1) + d_i(2)y_i(2) + \ldots + d_i(Z)y_i(Z)$

Only one of the Z possible outcomes is observed for each individual in the sample. We observe a propensity score, defined as $p_i(t = z \mid \mathbf{x}_i)$ and a propensity score vector

$\mathbf{p_i}(z, \mathbf{x_i}) = \{p_i(t = 1 \mid \mathbf{x}_i),\ p_i(t = 2 \mid \mathbf{x}_i), \ldots p_i(t = Z \mid \mathbf{x}_i)\}$, for each unit i.

Estimands:

$\forall\ z \neq z'$,

$$\text{ATE}_{z,\,z'} = \frac{\sum_{i=1}^{N} y_i d_i(z) w_{i,ATE}}{\sum_{i=1}^{N} d_i(z) w_{i,ATE}} - \frac{\sum_{i=1}^{N} y_i d_i(z\prime) w_{i,ATE}}{\sum_{i=1}^{N} d_i(z\prime) w_{i,ATE}}$$

$$\text{ATT}_{z,\,z'\,|\,z} = \frac{\sum_{i=1}^{N} y_i d_i(z) w_{i,ATT}}{\sum_{i=1}^{N} d_i(z) w_{i,ATT}} - \frac{\sum_{i=1}^{N} y_i d_i(z\prime) w_{i,ATT}}{\sum_{i=1}^{N} d_i(z\prime) w_{i,ATT}}$$

$$\text{ATU}_{z,\,z'\,|\,z'} = \frac{\sum_{i=1}^{N} y_i d_i(z) w_{i,ATU}}{\sum_{i=1}^{N} d_i(z) w_{i,ATU}} - \frac{\sum_{i=1}^{N} y_i d_i(z\prime) w_{i,ATU}}{\sum_{i=1}^{N} d_i(z\prime) w_{i,ATU}}$$

Weights:

Define j as an index of observations in treatment group z', where $j = \{1, 2, \ldots, N_{z'}\}$.

Define l as an index of observations in treatment group z, where $l = \{1, 2, \ldots, N_z\}$.

Inverse Probability of Treatment Weights (IPTW):

$$w_{i,ATE} = \begin{cases} \dfrac{1}{p_i(t = z\,|\mathbf{x}_i)}, & \forall\ i = l \\ \dfrac{1}{p_i(t = z'|\mathbf{x}_i)}, & \forall\ i = j \end{cases}$$

$$w_{i,ATT} = \begin{cases} 1, & \forall\ i = l \\ \dfrac{p_i(t = z\,|\mathbf{x}_i)}{p_i(t = z'\,|\mathbf{x}_i)}, & \forall\ i = j \end{cases}$$

$$w_{i,ATU} = \begin{cases} 1, & \forall\ i = j \\ \dfrac{p_i(t = z'|\mathbf{x}_i)}{p_i(t = z\,|\mathbf{x}_i)}, & \forall\ i = l \end{cases}$$

Kernel Weights (KW):

$$w_{i,ATT} = \begin{cases} 1, & \forall\ i = l \\ k_i(D_{lz}), & \forall\ i = j \end{cases}$$

$$k_i(D_{lz}) = \begin{cases} \dfrac{3}{4}\left(1 - \left(\dfrac{D_{lz}}{h}\right)^2\right), & if\ D_{lz} < h \\ 0, & otherwise \end{cases}$$

$D_{lz} = |\, p_i(t = z\,|\mathbf{x_i}) - p_l(t = z\,|\,\mathbf{x_i})\,|$

$$w_{i,ATU} = \begin{cases} 1, & \forall\ i = j \\ k_l\left(D_{jz'}\right), & \forall\ i = l \end{cases}$$

$$k_l\left(D_{jz'}\right) = \begin{cases} \dfrac{3}{4}\left(1 - \left(\dfrac{D_{jz'}}{h}\right)^2\right), & if\ D_{jz'} < h \\ 0, & otherwise \end{cases}$$

$D_{jz'} = |\, p_i(t = z'\,|\,\mathbf{x_i}) - p_j(t = z'\,|\,\mathbf{x_i})\,|$

$w_{i,\,ATE} = w_{i,ATT} + w_{i,\,ATU}$

Vector-Based Kernel Weighting (VBKW):

$$w_{i,ATT} = \begin{cases} 1, & \forall \ \ i = l \\ k_i(D_{lz}), & \forall \ \ i = j \end{cases}$$

$$k_i(D_{lz}) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{D_{lz}}{h}\right)^2\right), & if\ D_{lz} < h\ and\ D_{lm} \\ 0, & otherwise \end{cases}$$

$D_{lz} = | p_i(t = z \mid \mathbf{x_i}) - p_l(t = z \mid \mathbf{x_i}) |$

$D_{lm} = | p_i(t = m \mid \mathbf{x_i}) - p_l(t = m \mid \mathbf{x_i}) | \quad \forall\ \mathrm{m} \neq \mathrm{z}$

$$w_{i,ATU} = \begin{cases} 1, & \forall \ \ i = j \\ k_i\left(D_{jz'}\right), & \forall \ \ i = l \end{cases}$$

$$k_i\left(D_{jz'}\right) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{D_{jz'}}{h}\right)^2\right), & if\ D_{jz} < h\ and\ D_{jn} < h \\ 0, & otherwise \end{cases}$$

$D_{jz'} = | p_i(t = z' \mid \mathbf{x_i}) - p_j(t = z' \mid \mathbf{x_i}) |$

$D_{jn} = | p_i(t = n \mid \mathbf{x_i}) - p_j(t = n \mid \mathbf{x_i}) | \quad \forall\ \mathrm{n} \neq \mathrm{z'}$

$w_{i,\ ATE} = w_{i,ATT} + w_{i,\ ATU}$

Vector Matching (VM):

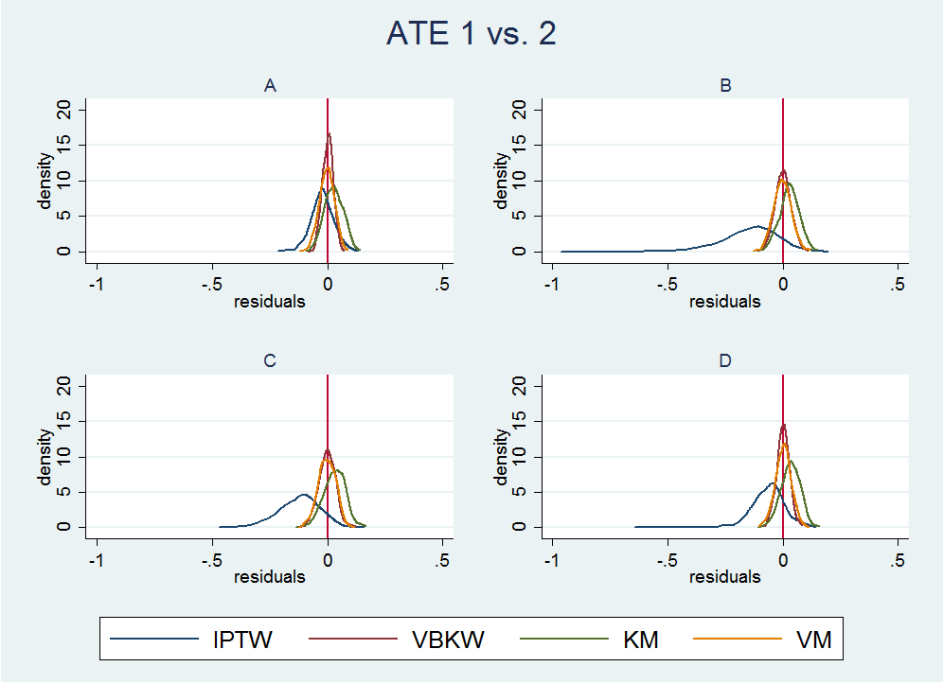
$w_{i,ATT} = n_{i,matched}$,where $n_{i,\ matched}$ is the number of times subject *i* is part of a matched set of observations composed of at least one individual from each treatment group, when matching is implemented using reference group z.

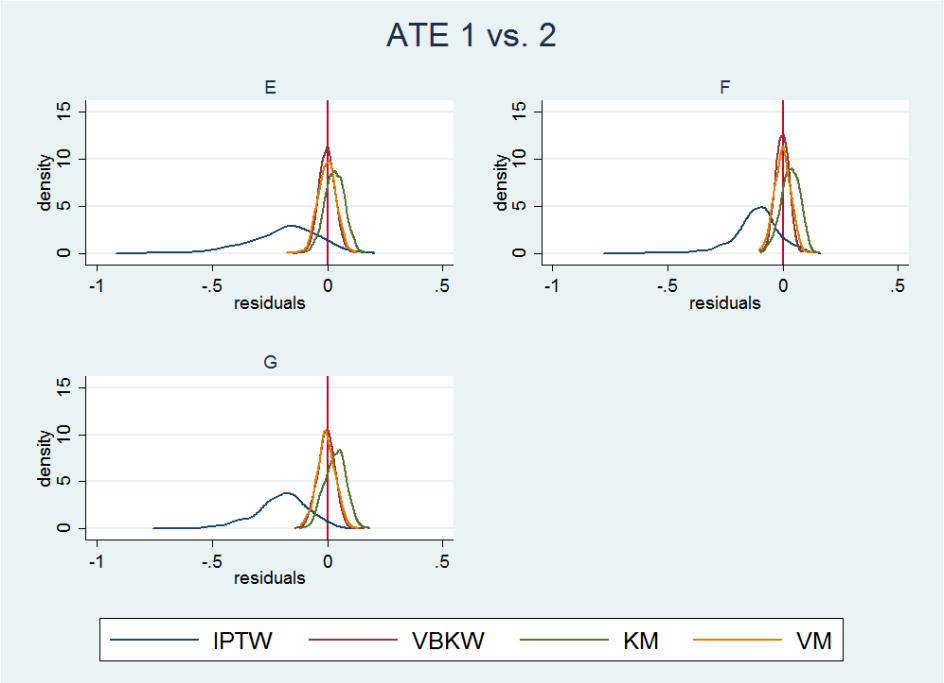
$w_{i,ATU} = n_{i,matched}$,where $n_{i,\ matched}$ is the number of times subject *i* is part of a matched set of observations composed of at least one individual from each treatment group when matching is implemented using reference group z'.

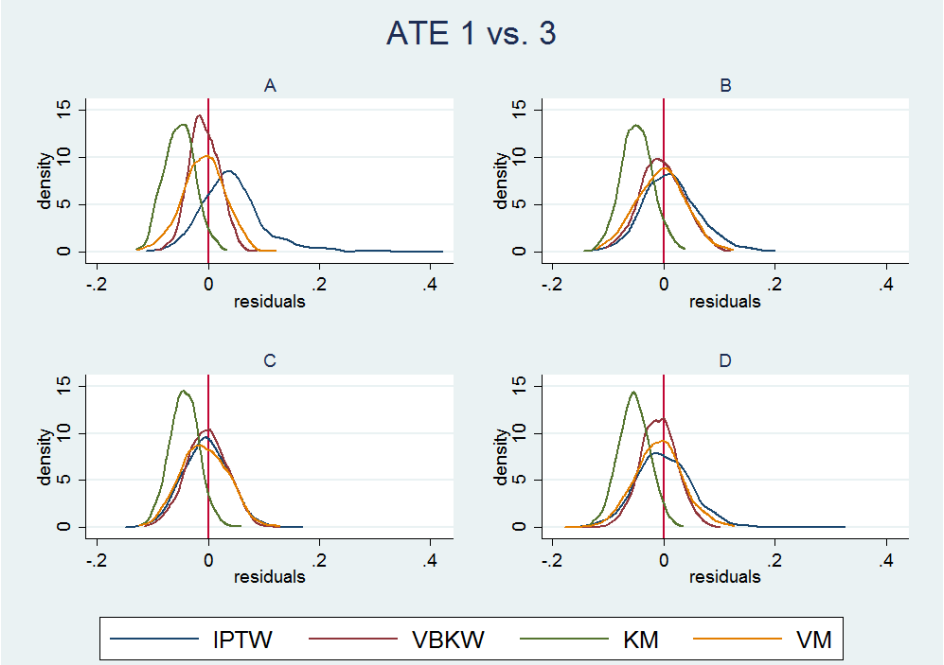
$w_{i,\ ATE} = w_{i,ATT} + w_{i,\ ATU}$

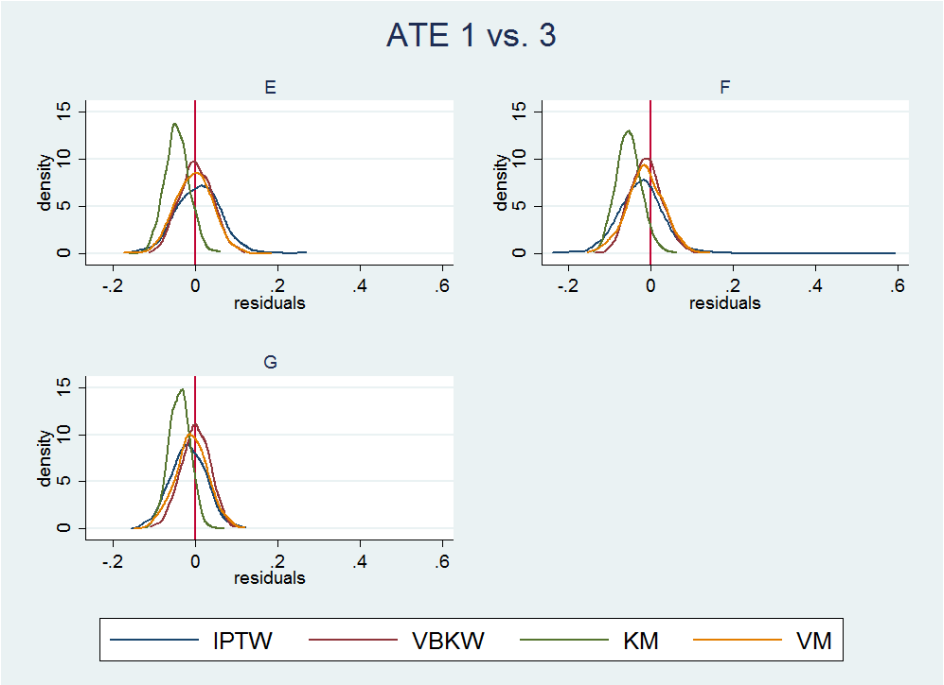
Appendix 2

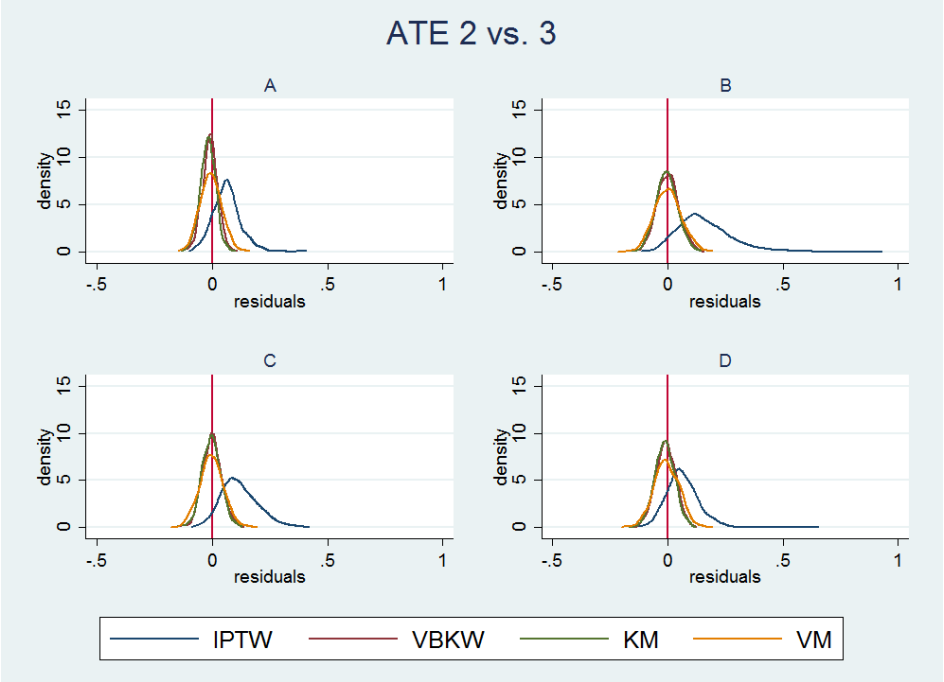
Density plots of residuals from ATE and ATT estimates using IPTW, KW, VM, and VBKW, even distribution of sample across three treatment groups, and n=999. Estimates for homogeneous treatment effects are presented. True propensity score includes A) main effects only, B) mild nonlinearity, C) moderate nonlinearity, D) mild non-additivity, E) mild nonlinearity and mild non-additivity, F) moderate non-additivity, G) moderate nonlinearity and moderate non-additivity (Setoguchi et al. 2008). IPTW = inverse probability of treatment weights, KM = kernel weights, VBKW = vector-based kernel weights, VM = vector matching.

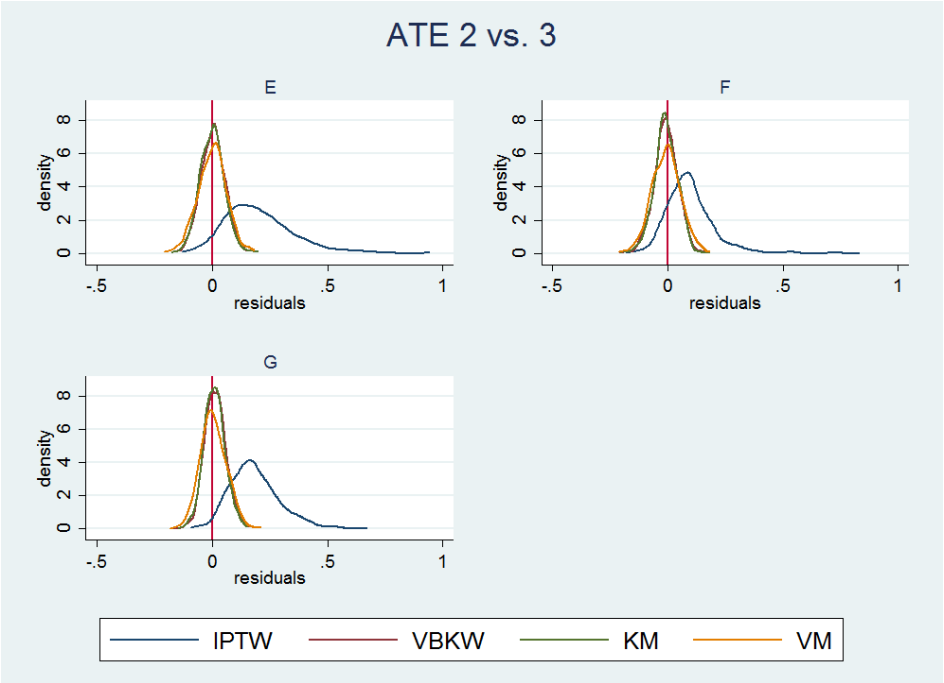

ATE 1 vs. 2
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

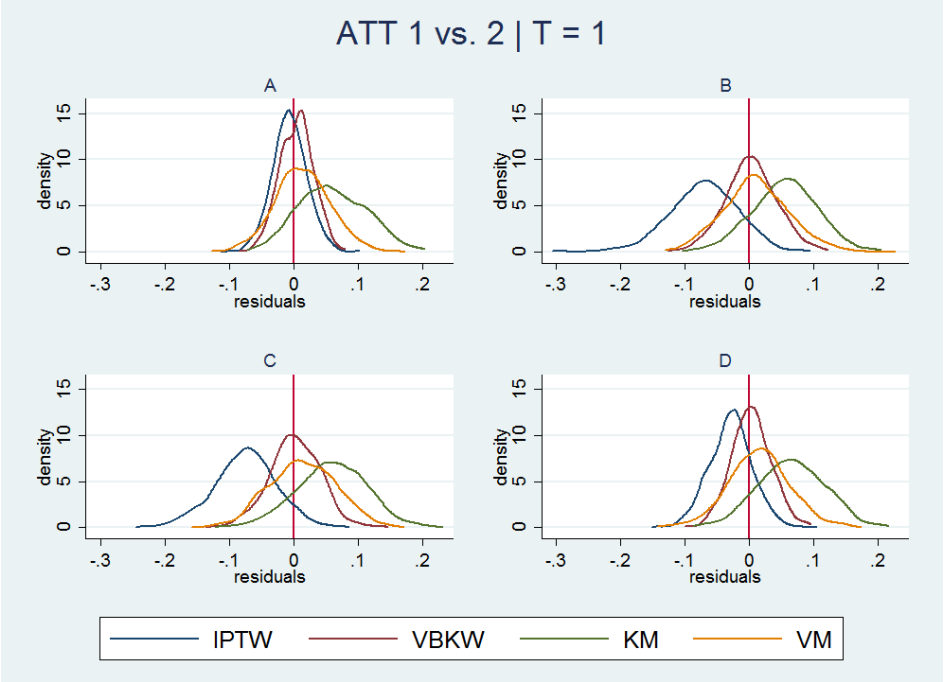

ATE 1 vs. 2
E
F
G
density
residuals
IPTW
VBKW
KM
VM

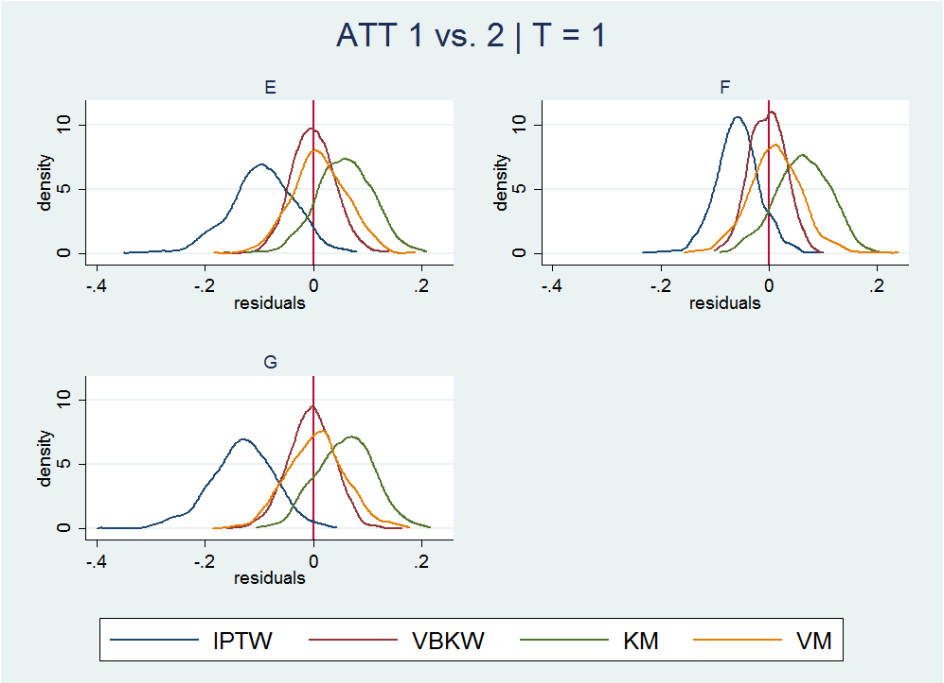

ATE 1 vs. 3
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

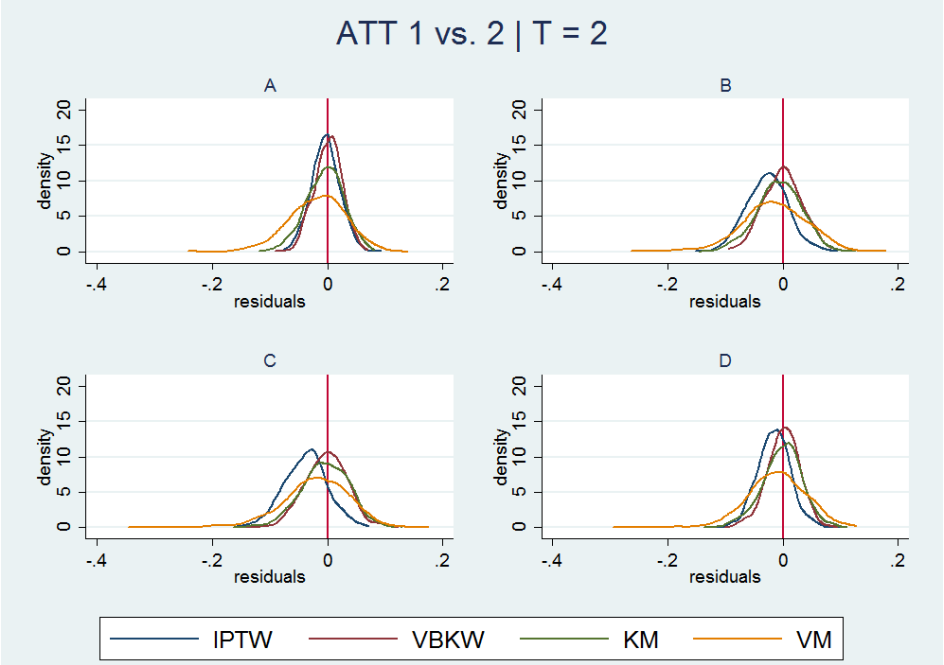

ATE 1 vs. 3
E
F
G
density
residuals
IPTW
VBKW
KM
VM

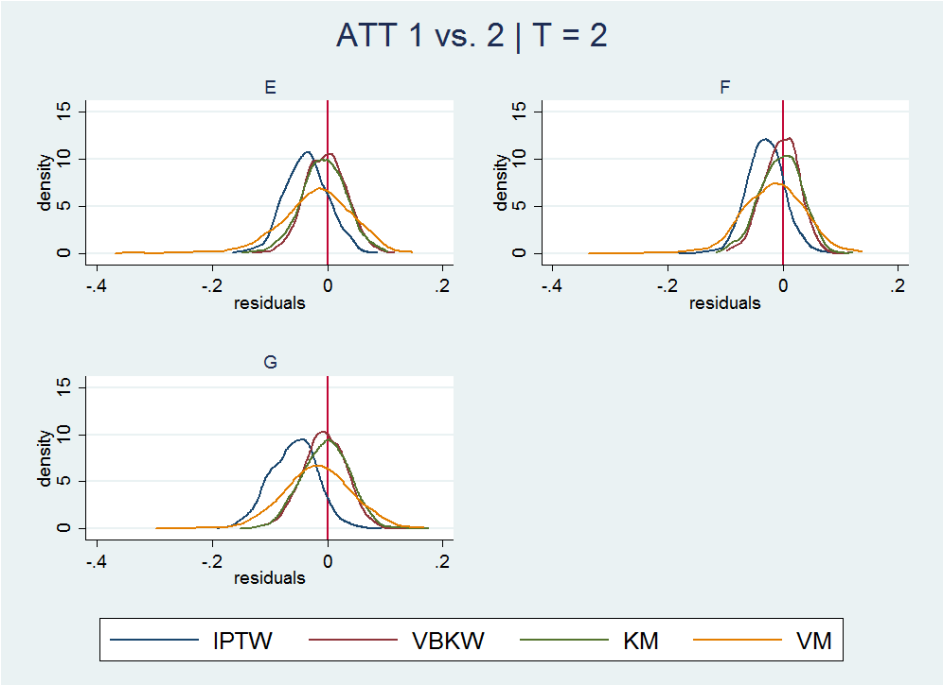

ATE 2 vs. 3
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

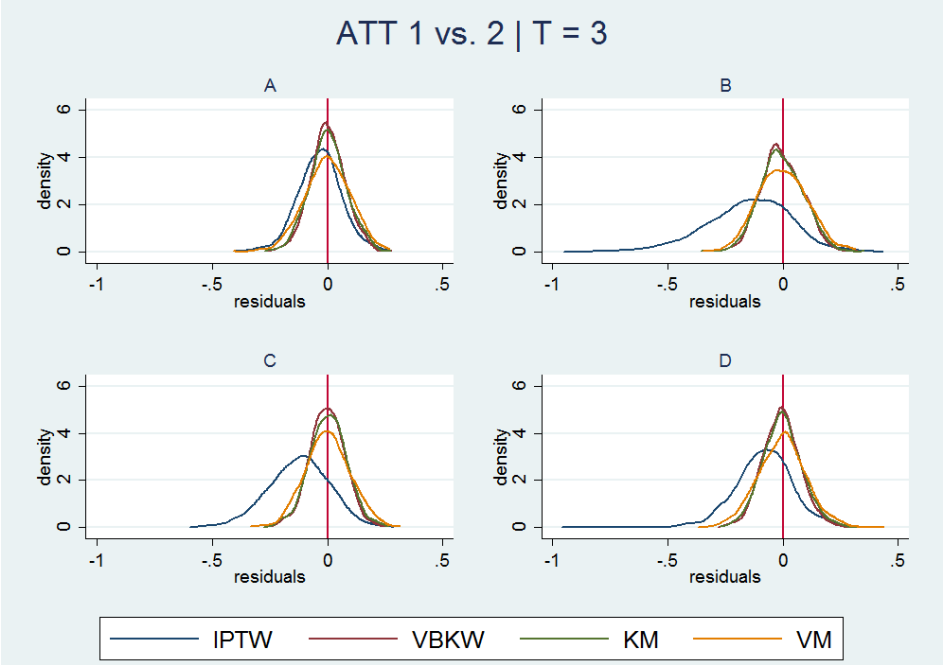

ATE 2 vs. 3
E
F
G
density
residuals
IPTW
VBKW
KM
VM

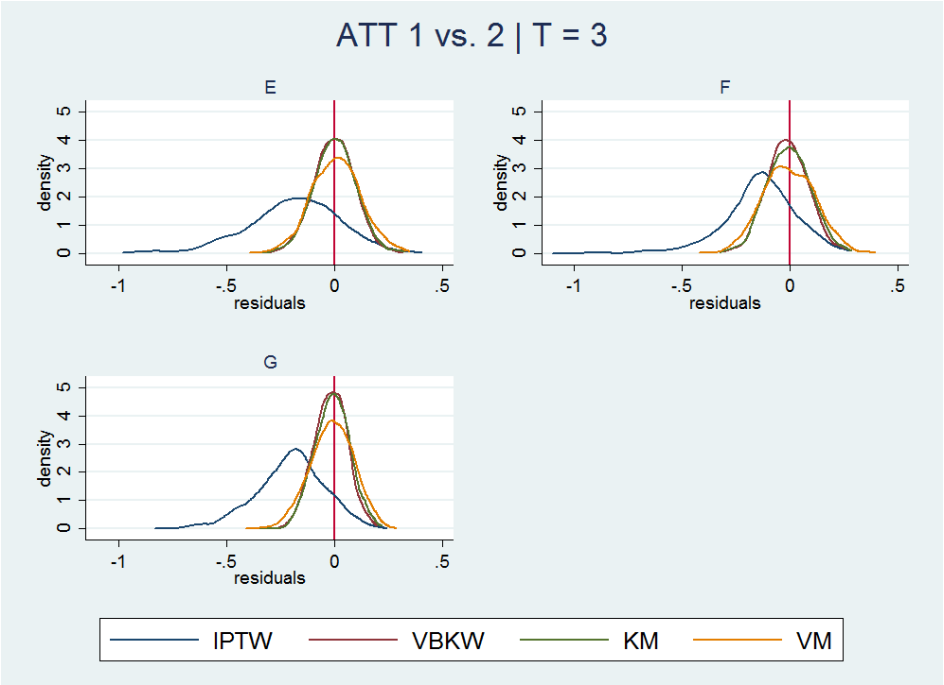

ATT 1 vs. 2 | T = 1
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

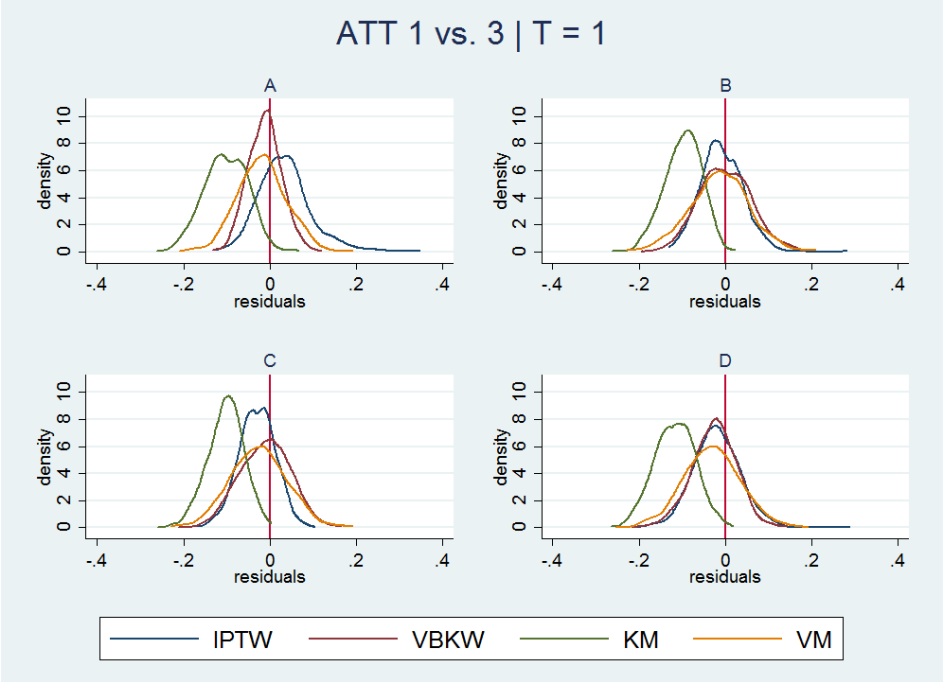

ATT 1 vs. 2 | T = 1
E
F
G
density
residuals
IPTW
VBKW
KM
VM

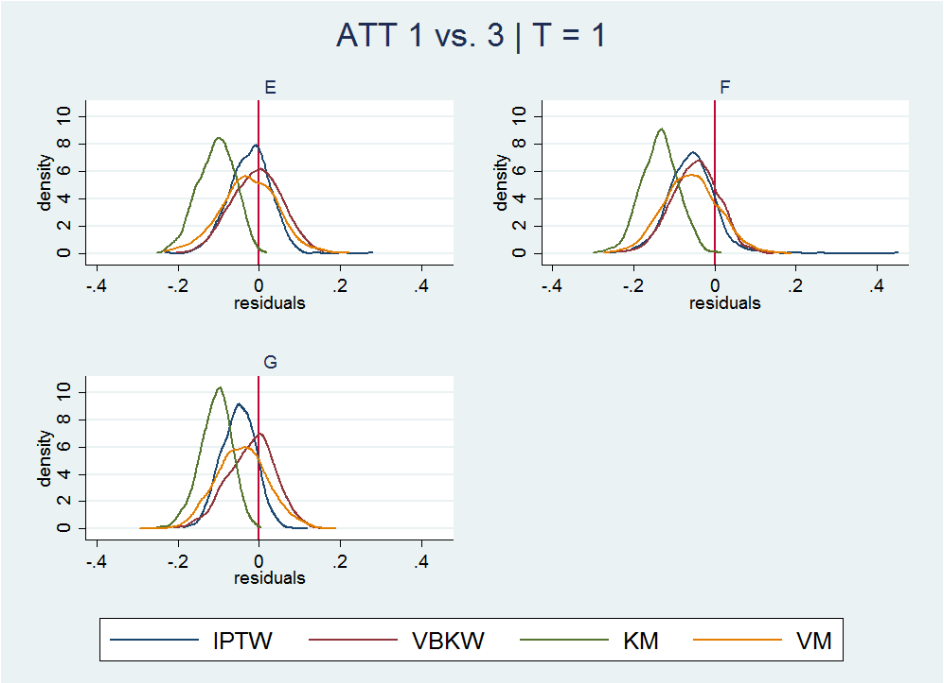

ATT 1 vs. 2 | T = 2
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

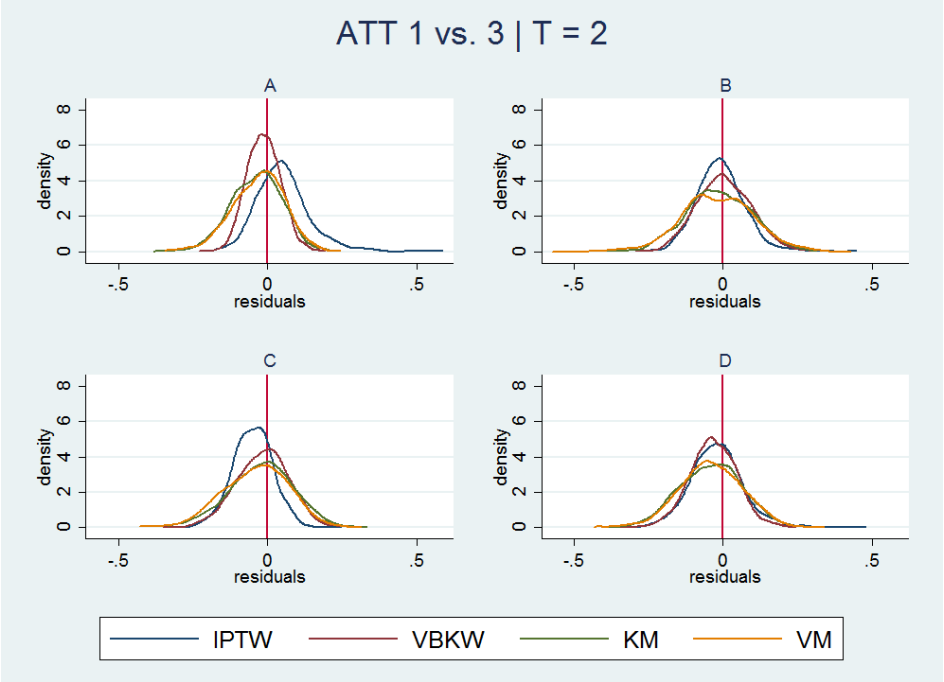

ATT 1 vs. 2 | T = 2
E
F
G
density
residuals
IPTW
VBKW
KM
VM

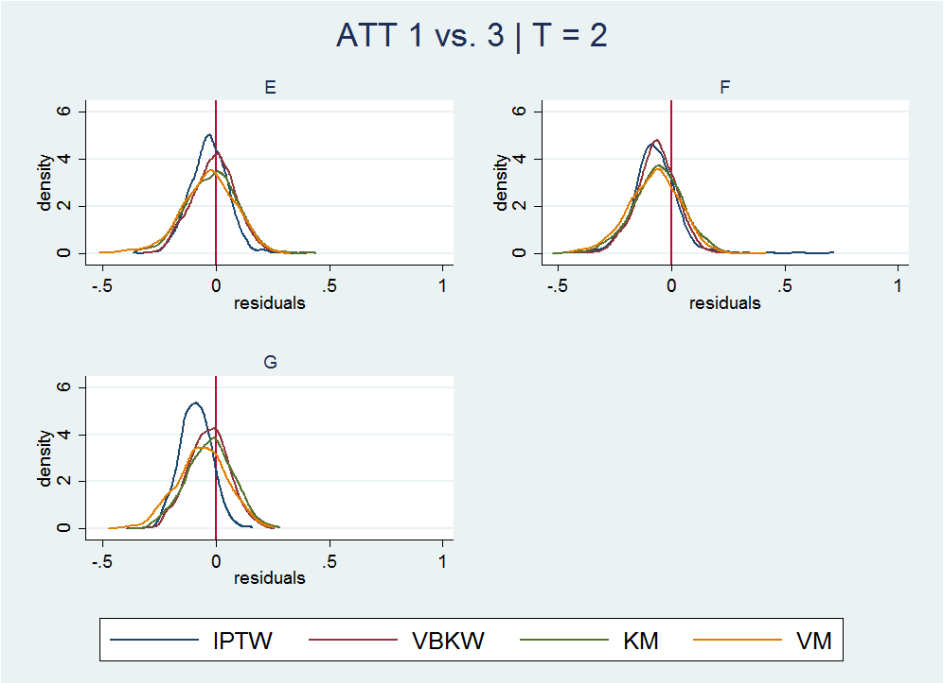

ATT 1 vs. 2 | T = 3
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

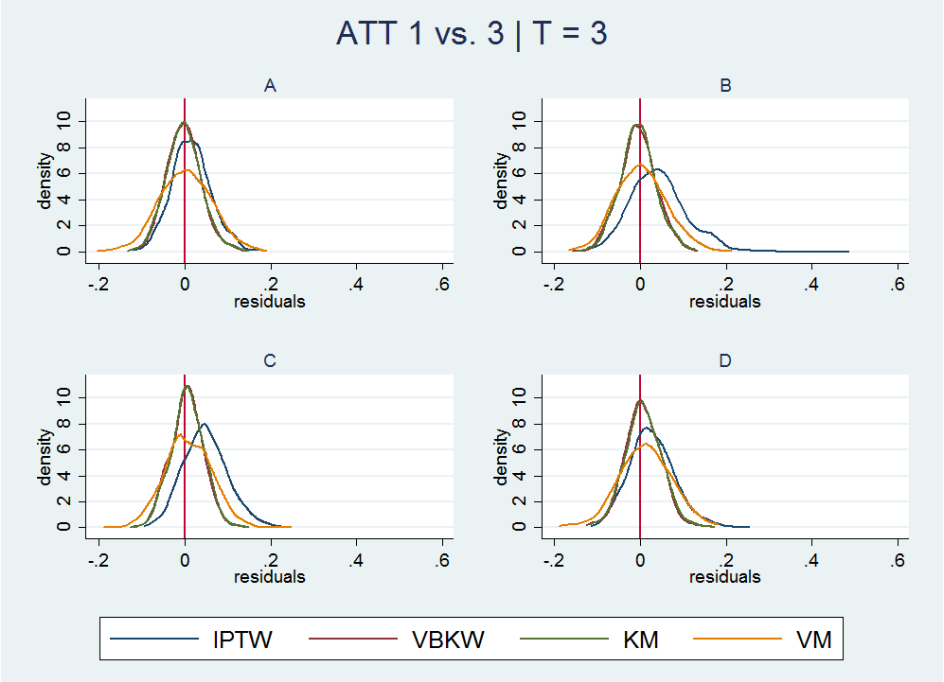

ATT 1 vs. 2 | T = 3
E
F
G
density
residuals
IPTW
VBKW
KM
VM

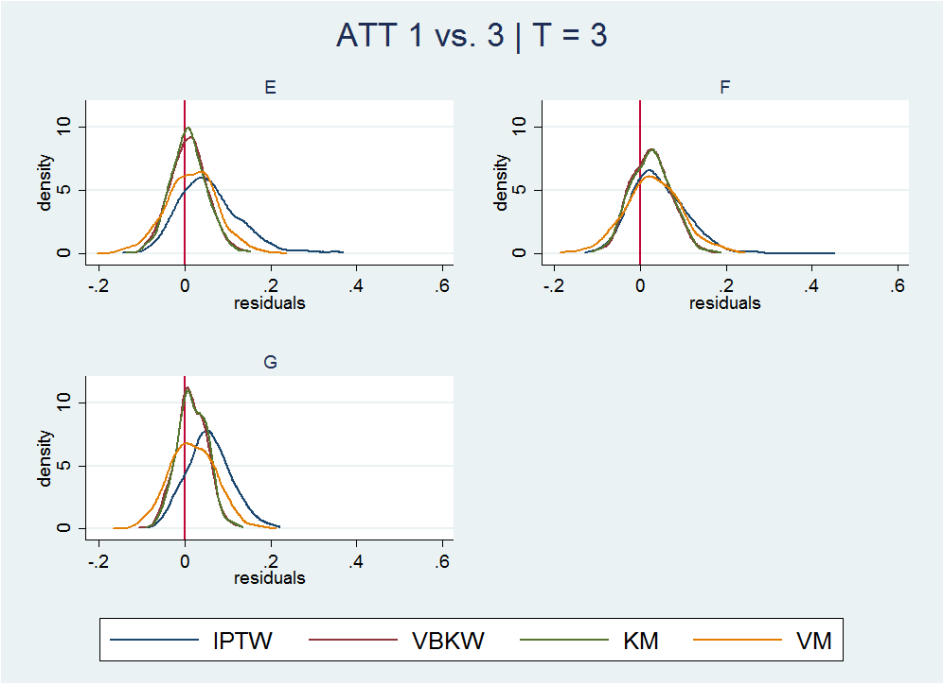

ATT 1 vs. 3 | T = 1
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

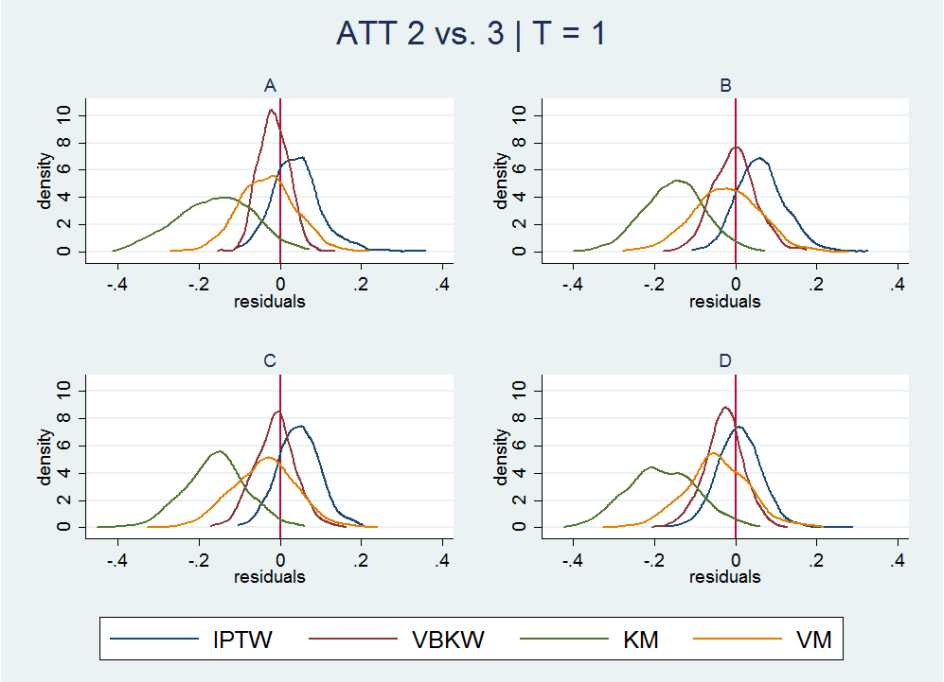

ATT 1 vs. 3 | T = 1
E
F
G
density
residuals
IPTW
VBKW
KM
VM

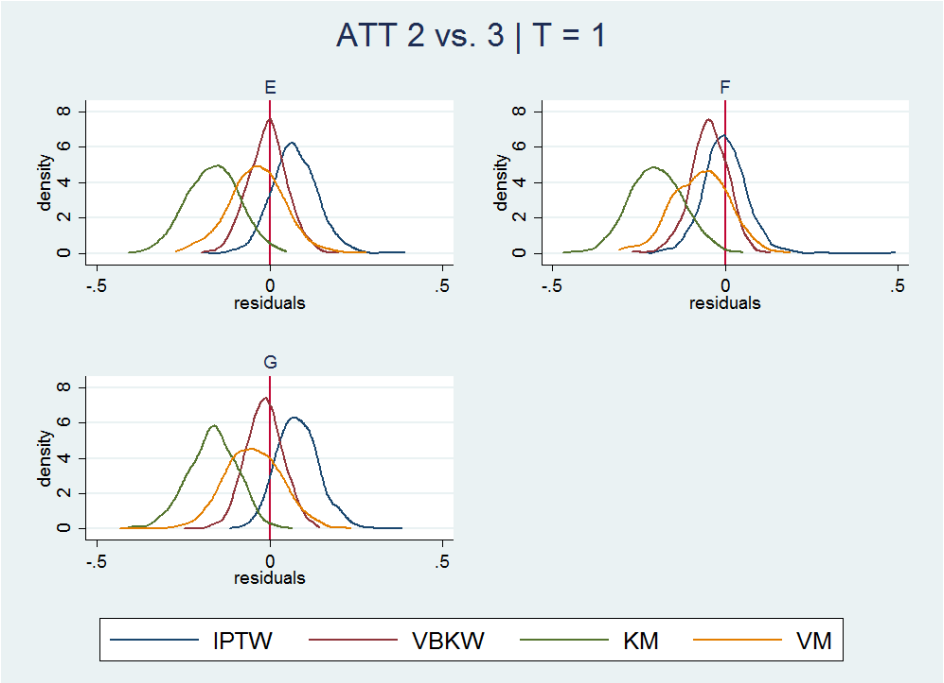

ATT 1 vs. 3 | T = 2
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

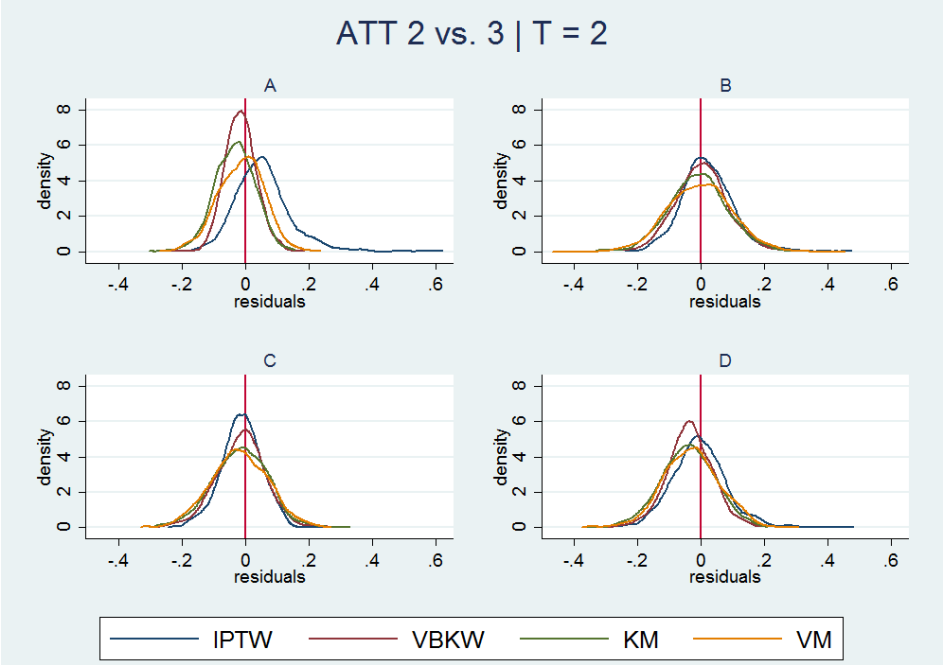

ATT 1 vs. 3 | T = 2
E
F
G
density
residuals
IPTW
VBKW
KM
VM

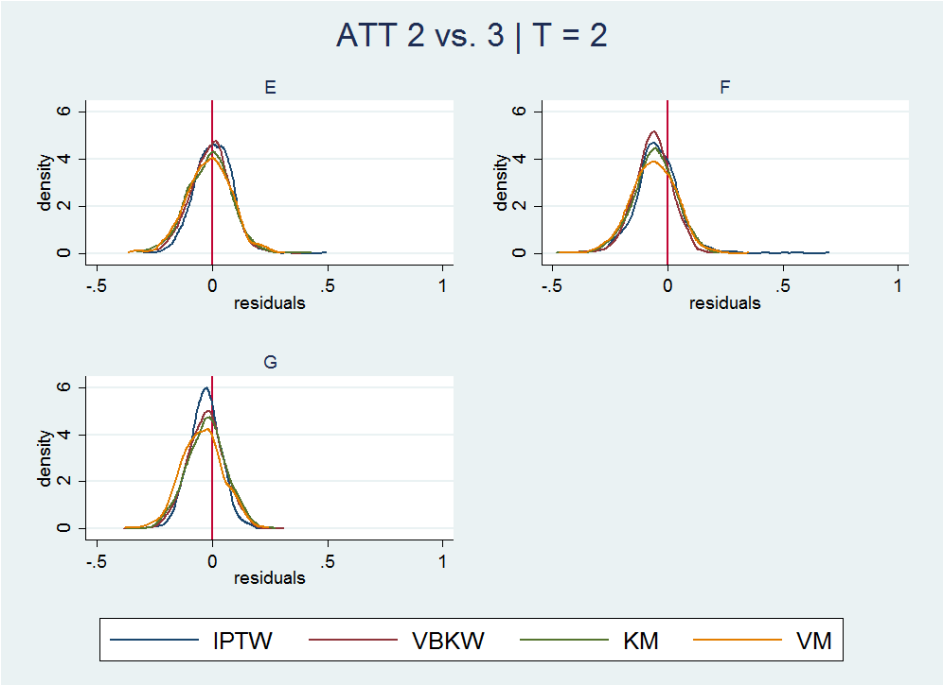

ATT 1 vs. 3 | T = 3
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM

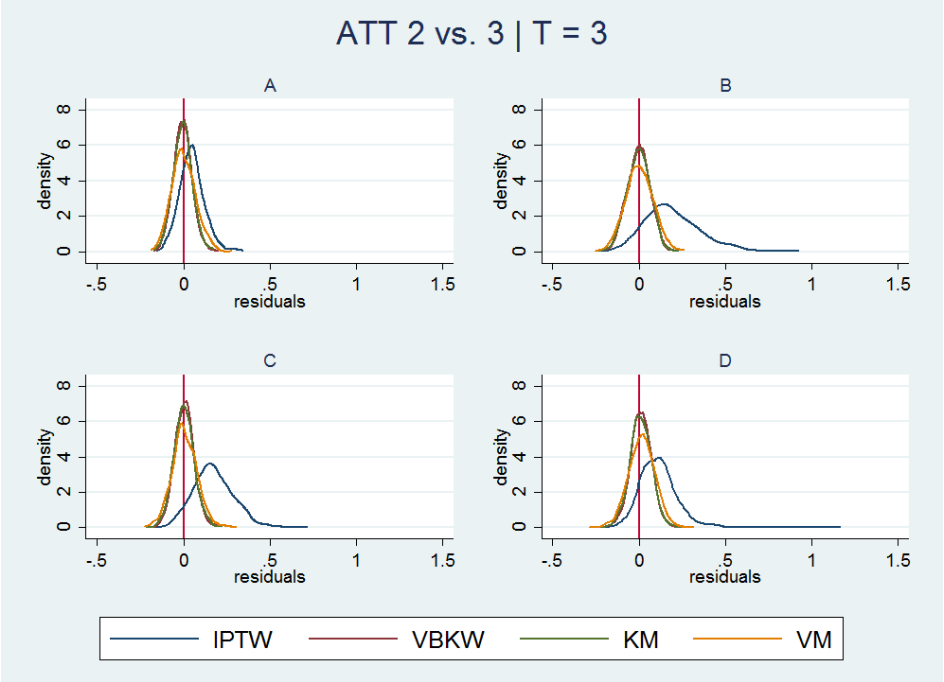

ATT 1 vs. 3 | T = 3
E
F
G
density
residuals
IPTW
VBKW
KM
VM

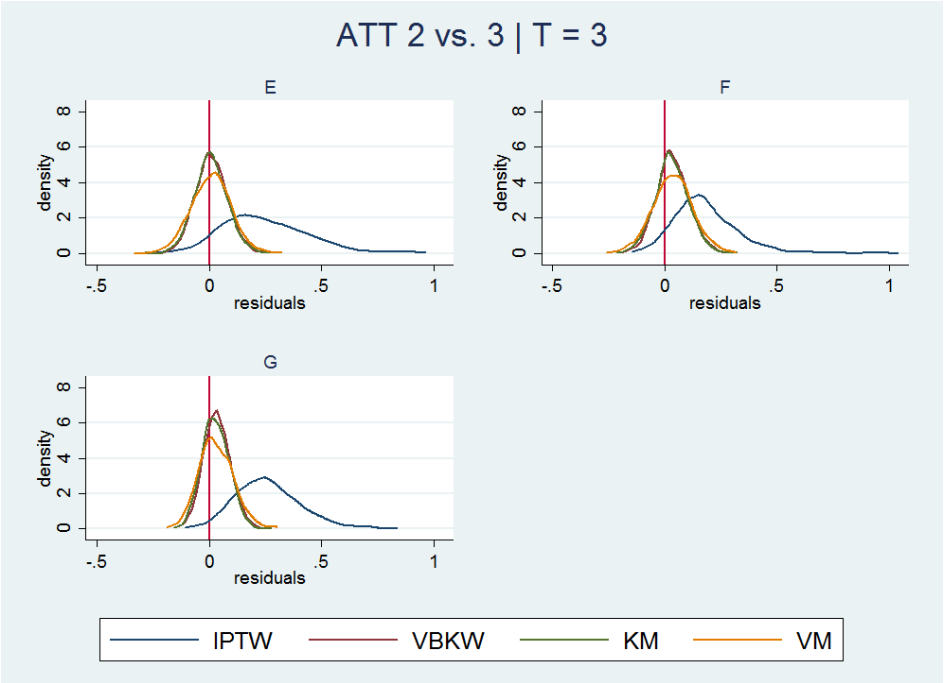

ATT 2 vs. 3 | T = 1
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM


ATT 2 vs. 3 | T = 1
E
F
G
density
residuals
IPTW
VBKW
KM
VM


ATT 2 vs. 3 | T = 2
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM


ATT 2 vs. 3 | T = 2
E
F
G
density
residuals
IPTW
VBKW
KM
VM


ATT 2 vs. 3 | T = 3
A
B
C
D
density
residuals
IPTW
VBKW
KM
VM


ATT 2 vs. 3 | T = 3
E
F
G
density
residuals
IPTW
VBKW
KM
VM